DOCUMENT RESUME

ED 243 897 TM 830 527

AUTHOR	Swinton, Spencer S.
TITLE	A Manual for Assessing Language Growth in Instructional Settings.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-83-17; TOEFL-RR-14
PUB DATE	Feb 83
NOTE	200p.
AVAILABLE FROM	Educational Testing Service, Research Publications R-116, Princeton, NJ 08541.
PUB TYPE	Guides - Non-Classroom Use (055) -- Reports - Research/Technical (143)
EDRS PRICE	MF01/PC08 Plus Postage.
DESCRIPTORS	*Achievement Gains; Computer Software; Data Analysis; Data Collection; *English (Second Language); Glossaries; Higher Education; Language Tests; Predictive Measurement; *Pretests Posttests; *Regression (Statistics); Scores; *Test Reliability
IDENTIFIERS	Statistical Package for the Social Sciences; *Test of English as a Foreign Language

ABSTRACT

This manual is designed to assist administrators of English-as-a-second-language programs in assessing students' language growth. It begins by reviewing some of the concepts and terminology to be used. It then goes on to suggest and illustrate data-recording formats and methods of summarizing raw gains. This is followed by an example based on bowling scores to illustrate the regression effect. An overview of a method for separating raw gain into regression and true gain components follows. It concludes with a brief discussion of a method for comparing two different groups with differing backgrounds or curricula. The appendices give details of the data and of the steps in performing the regression analyses using SPSS (Statistical Package for the Social Sciences). (BW)

Research Reports

REPORT 14
FEBRUARY 1983

A MANUAL FOR ASSESSING LANGUAGE GROWTH IN INSTRUCTIONAL SETTINGS

Spencer S. Swinton

EDUCATIONAL TESTING SERVICE

The Test of English as a Foreign Language (TOEFL) was developed in 1963 by a National Council on the Testing of English as a Foreign Language, which was formed through the cooperative effort of over thirty organizations, public and private, that were concerned with testing the English proficiency of non-native speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service (ETS) and the College Board assumed joint responsibility for the program and in 1973 a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations (GRE) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education.

ETS administers the TOEFL program under the general direction of a Policy Council that was established by, and is affiliated with, the sponsoring organizations. Members of the Policy Council represent the College Board and the GRE Board and such institutions and agencies as graduate schools of business, junior and community colleges, nonprofit educational exchange agencies, and agencies of the United States government.

A continuing program of research related to TOEFL is carried out under the direction of the TOEFL Research Committee. Its six members include representatives of the Policy Council, the TOEFL Committee of Examiners, and distinguished English-as-a-second-language specialists from the academic community. Currently the committee meets twice yearly to review and approve proposals for test-related research and to set guidelines for the entire scope of the TOEFL research program. Members of the Research Committee serve three-year terms at the invitation of the Policy Council; the chair of the committee serves on the Policy Council.

Because the studies are specific to the test and the testing program, most of the actual research is conducted by ETS staff rather than by outside researchers. However, many projects require the cooperation of other institutions, particularly those with programs in the teaching of English as a foreign or second language. Representatives of such programs who are interested in participating in or conducting TOEFL-related research are invited to contact the TOEFL program office. Local research may sometimes require access to TOEFL data. In such cases, the program may provide this data following approval by the Research Committee. All TOEFL research projects must undergo appropriate ETS review to ascertain that the confidentiality of data will be protected.

Current (1981-82) members of the TOEFL Research Committee include the following:

G. Richard Tucker (chair)	Center for Applied Linguistics
Louis A. Arena	University of Delaware
H. Douglas Brown	University of Illinois at Urbana-Champagne
Frances B. Hinofotis	University of California at Los Angeles
Diane Larsen-Freeman	The Experiment in International Living
David S. Sparks	University of Maryland

A Manual for Assessing Language Growth in Instructional Settings

Spencer S. Swinton

Educational Testing Service
Princeton, New Jersey

RR 83-17

Table of Contents

List of Tables

List of Figures

List of Appendices

Acknowledgments

The preparation of this manual would not have been possible without the generous assistance and suggestions of Donald L. Alderman, Paul J. Angelis, Louis A. Arena, Allis R. Bens, H. Douglas Brown, Rosalea G. Courtney, Dorothy Danielson, Frances B. Hinofotis, Vera L. Jones, Donald Knapp, Diane Larsen-Freeman, Donald E. Powers, Allen W. Sharp, David S. Sparks, Charles W. Stansfield, G. Richard Tucker, Russell Webster, and Kenneth M. Wilson. Any errors remain the responsibility of the author.

Introduction

A number of United States colleges and universities provide intensive English-as-a-second-language (ESL) instruction for entering foreign students. Generally, these programs consist of full-time English language instruction over a period of several semesters. Students typically graduate to English-medium instruction in their chosen fields, in the same or another institution, only after meeting some entrance criterion score on an English language proficiency examination, such as the Test of English as a Foreign Language (TOEFL) or the Michigan Test of English Language Proficiency.

This manual is designed to assist administrators of ESL programs in assessing students' language growth. It is a guide for conducting studies at local institutions to predict likely progress over time of students at different entry levels. It also offers assistance in interpreting the typical outcome; that is, students who enter with low scores frequently show greater gains than do students who enter with higher scores. This last issue is linked to the classical "regression to the mean" phenomenon, which occurs with any test that is not perfectly reliable. We suggest a method for assessing the amount to which lower-scoring students can be expected to show relatively more apparent gain due to unreliability in measurement. If the low scorer's "edge" in your program is in the neighborhood of this expected value as calculated from your students' scores, there is no evidence from the scores that the curriculum at the more advanced level is not producing adequate results. (There could, of course, be other bases for such a change in emphasis, but, as shall be demonstrated, even average raw score gains of 80 points for students who start with TOEFL scores of 300, versus raw gains of 30 points for students who begin with scores of 500, do not necessarily mean that the curriculum is less effective for the higher proficiency group.)

The manual begins by reviewing some of the concepts and terminology to be used. It then goes on to suggest and illustrate data-recording formats and methods of summarizing raw gains. This is followed by an example based on bowling scores to illustrate the regression effect. An overview of a method for separating raw gain into regression and true gain components follows. It concludes with a brief discussion of a method for comparing two different groups with differing backgrounds or curricula. A summary of the steps in the recommended analysis is given in Figure 1.

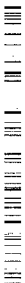
The appendices give details of the data and of the steps in performing the regression analyses using SPSS (Statistical Package for the Social Sciences), a widely available statistical analysis computer package. If SPSS is not available at your computer installation, students or staff should be able to adapt this sample to other regression programs with little difficulty. By working through the example and discussion beginning on page 5, the reader should understand what these methods can do, and be in a position to decide whether to forward the appendices to a staff member or student familiar with SPSS to conduct analyses of local data. If such analyses are performed, the last few pages of this section

Figure 1

Steps in Analysis

Beginning of Semester:	1.	Administer Pretest (A)
	2.	Score Pretest (A)
	3.	Record Data
One Week Later:	4.	Administer Reliability Test (B)
	5.	Score Reliability Test (B)
	6.	Record Data
End of Semester:	7.	Administer Posttest (C)
	8.	Score Posttest (C)
	9.	Record Data
Analysis:	10.	Keypunch Data
	11.	Match Scores
Estimate Baseline Equation:	12.	Predict Reliability Test from Pretest (Regression Equation)
Estimate Posttest Equation:	13.	Predict Posttest from Pretest (Regression Equation)
Estimate Growth:	14.	For each pretest score, A_i, estimate expected growth by subtracting predicted reliability test score, B_i, from predicted posttest score, C_i, obtaining $C_i - B_i$ rather than raw gain, $C_i - A_i$.



provide an overview of the interpretation of the analyses for the illustrative data used here. The appendices themselves--A on data layout, B on gain analyses, and C on comparative growth of two groups--give more detail for the use of individuals who perform the computer analyses.

This manual uses illustrative data collected from 98 ESL students at San Francisco State University. With the cooperation of faculty members and Allis R. Bens (director of the American Language Institute at the university), students were pretested with TOEFL before the fall semester began, administered a reliability test one week into the semester, and given a posttest at the end of the semester. These institutional TOEFL administrations were scored at ETS and the scores were returned to the university. The analyses reported here were performed at ETS.

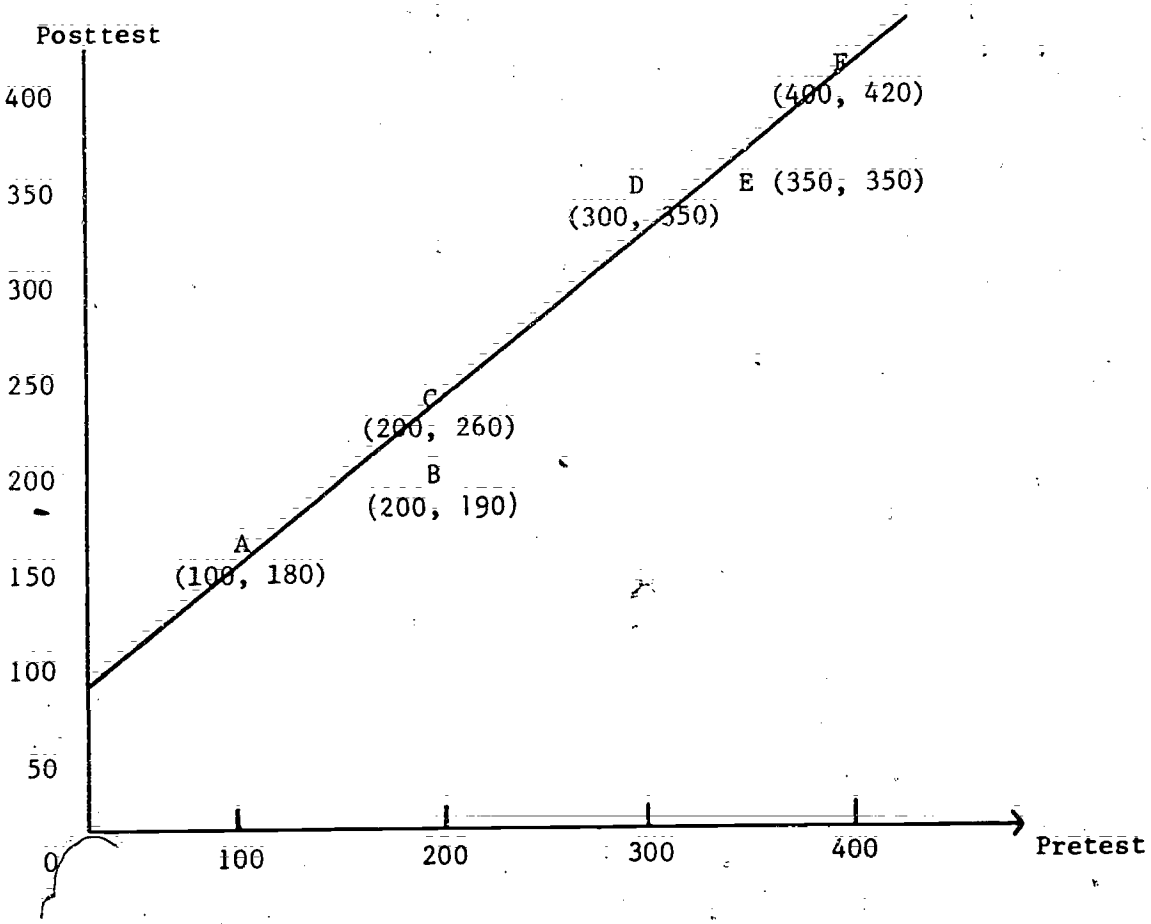
The major impetus for the development of this manual came from Professor Bens's gentle but firm insistence that her internal analyses of raw gain for varying pretest score ranges could be made more rigorous without becoming incomprehensible. The intent was to go beyond the general nostrum that "some regression to the mean is inevitable," and to find ways of estimating how much regression is to be expected for a given group, and how much of the apparent change may be attributed to language growth. Earlier versions of the manual were discussed with San Francisco State faculty members. The current version has been revised further on the basis of user reaction.

When we are confronted with the two sets of numbers, such as the pretest and posttest scores of a group of students, we must first summarize them if we wish to discern any general patterns and relationships that may be concealed under the blooming, buzzing confusion of their individual jumps and wiggles. The best way to detect and understand underlying patterns is by summarizing and simplifying a plot of the data. The most concise way to summarize and manipulate the data for statistical tests is to reduce important features of the plot to the simple linear equations of elementary analytic geometry.

Thus, we need to develop some facility in translating from plots to formulas and back again. Suppose we have the following pairs of pretest and posttest scores on six students:

Student	A	B	C	D	E	F
Pretest	100	200	200	300	350	400
Posttest	180	190	260	350	350	420

It is customary to plot such scores on a graph with pretest coordinates on the horizontal axis and posttest coordinates on the vertical axis, so that each student's pair of numbers determines a point on the grid. Following is a plot of the points determined by these pairs of numbers.



Clearly, these points lie very near an "uphill" straight line passing close to A and F. The "best-fitting" straight line for these six points could be defined in several ways, but if we wish to predict posttest from pretest, the best-fitting straight line is defined as the line that minimizes the sum of the squared vertical distances of the points from the line. If you place a clear straightedge over the points and line it up by eye to come as close as it can to have an equal number of evenly distributed points above and below the edge, you will probably come remarkably close to the mathematically defined "regression" line of best fit. Translating this line to a mathematical equation is straightforward. A line can be described by two constants: the slope and the intercept.

The line in the figure cuts the posttest scale at about 80, when extended to the vertical axis. This point corresponds to the theoretical value of the posttest score when the pretest is zero. The distance that the line cuts the vertical axis above or below the origin is called the intercept, and appears in the equation of the line as an additive constant. The line in this example can be described by the equation posttest = 80 + .85 (pretest).

That is, for any pretest score multiply by .85, add 80, and the corresponding vertical coordinate of the line will result. Clearly, when

the pretest score is 0, the posttest score is just 80. When the pretest score is 400, the equation gives a posttest score of .85 x 400 + 80, or 340 + 80 = 420. The multiplicative constant, in this case .85, tells how fast the line climbs as we move to the right. For every additional pretest point, this line climbs .85 points. For every hundred additional pretest points, the posttest climbs 85 points. The multiplicative constant is called the slope of the line. If the slope is 1, the line climbs at a 45° angle when pretest and posttest are plotted in the same units. If the slope is 0, the line is horizontal and there is no relationship between pretest and posttest scores. In the kinds of relationships that we will be examining, the slope is usually between 0 and 1, unless one or both tests are too easy or too difficult. In such cases, the relationship may not be a straight line, and will not be characterizable by a single value of slope across the range of interest. The simple models discussed in this manual will then not apply.

We turn now to a suggested format for recording a program's data, and for tabulating and summarizing scores manually to detect patterns of change.

Recording Data

Monitoring language growth serves several important functions related to the control of student flow through the program. First, an estimate of the likely scores at the end of the semester for students with various entering scores helps institutions plan for enrollment in courses at various levels in the following semester. A discussion of approximate methods for obtaining such estimates from locally derived score summaries begins on page 7.

Second, individual students can be helped to estimate the likely number of semesters of instruction they will need to achieve a particular level of proficiency. It must, of course, be remembered that there is considerable variation around any average level of performance.

A third use of score records is program evaluation. Is the sequence becoming more or less effective over the years? Is a new textbook series better than the current texts for the development of structure and written expression? Such comparisons are introduced on page 13 and discussed in detail in Appendix C. Does the course have as great an impact on intermediate English students as on beginners? This last question is complicated by the "regression effect," and was a major impetus for the development of this manual. The question is introduced on page 14 and followed by an outlined solution on pages 15-19. A summary of the results of applying a computer program to illustrative data begins on page 19. Details of running the computer regression package are provided in Appendix B.

By collecting data in a form suitable for easy retrieval and analysis, and by using some of the techniques suggested in this manual,

Table 1

Continuous Progress Form

(Prototype)

	1	2	3	4			5		6			7		8			9		10
NAME	NATIVE LANGUAGE	DATE ENTERED PROGRAM	YEARS OF PRIOR ENGLISH	TOEFL SCORES LC	SW	RV	TOTAL	DATE	TOEFL SCORES LC	SW	RV	TOTAL	DATE	TOEFL SCORES LC	SW	RV	TOTAL	DATE	SEM. GRADE
CONSUELO M.T.	SPANISH	9/4/80	4	43	32	32	357	9/8/80	42	38	31	370	9/15/80	50	41	33	413	1/30/81	B
2				50	41	33	413	1/30/81	49	41	35	417	2/16/81	54	45	42	470	6/19/81	A
3																			
4																			
HUSSEIN, A.B. 5	ARABIC	2/5/81	6	48	46	46	467	2/9/81	48	49	47	480	2/16/81	52	52	48	507	6/19/81	B
6																			
7																			
8																			
9																			
10																			
11																			
12																			
13																			
14																			
15																			
16																			
17																			
18																			
19																			
21																			

such questions may be addressed with data from your institution. The use of uniform data collection and reporting formats also facilitates pooling of data across years, programs, and institutions. This makes it possible to study such questions as the impact of various teaching approaches on students with uncommon language backgrounds, who might not appear in sufficiently large numbers in any one institution to make such a study possible.

One useful format for record keeping is a continuous progress form, with student identification and background information, instructional history, grades, and test scores over the student's ESL career. An example is given in Table 1. Four rows have been allotted to each student. Programs that typically retain students for more semesters could adjust the form accordingly. Space has been provided to record subscores and totals of three institutional administrations of TOEFL each semester. The second of the three TOEFL testings, given one week after the pretest, is for the purpose of establishing a baseline for growth studies, and will be discussed in the next section. Additional columns and the reverse of the form could be used for instructor and program information and additional student data. It is easy to keypunch scores from such records directly onto cards or tape, as in Appendix A.

The following paragraphs discuss ways of summarizing such data without a computer to make it easier to calculate averages and to develop simple graphs that may clarify relationships between pretest and posttest. To make the discussion concrete, we use TOEFL scores collected from a group of students in one semester of an intensive ESL course.

Summarizing Data

A group of 98 students in a full-time intensive English course at San Francisco State University took an institutional administration of TOEFL as a pretest prior to, or at the beginning of, the fall semester. One week after the beginning of the term, they took another form of the test (about which more later). At the end of the 13-week semester, they were administered a third form of TOEFL as a posttest.

The means and standard deviations of the pretest and posttest are given in Table 2.

Table 2

Means and Standard Deviations,
San Francisco State Gain Study

	Mean	Standard Deviation
Pretest	399.77	61.57
Posttest	455.34	59.94

Program administrators are likely to want to know if a gain of about 56 points can be expected at all points on the TOEFL scale, if those with lower pretest scores are likely to grow relatively more, or if those who start with higher pretest scores can be expected to grow relatively more.

One way of examining this issue is to group posttest scores by range of pretest score, as in Table 3. (The complete set of scores and subscores are given in Appendix A.) The first individual's pre- and postscores are 473 and 527, so 473 is in the pretest range 451-500. In Table 3, 527 is entered at (a), the first posttest score in the pretest range column 451-500. The second student's scores are 357 and 403. The posttest score, 403, is thus the first entry at (b) in the pretest range column 351-400. Once the 98 posttest scores have been entered, each column total is divided by the number in that column to obtain the column average. The average of the pretest scores for the column has also been calculated from a similar table and entered at the top of each column. Table 4 summarizes these averages.

Table 3

Posttest Scores for Various Pretest Ranges

(Prototype)

Pretest Mean	293.5	327.8	379.6	426.2	469.5	523.5	557.0	613.0	
Pre-test Range	251-300	301-350	351-400	401-450	451-500	501-550	551-600	601-650	
Posttest n	4	19	29	27	15	2	1	1	=98
	420	370	(b) 403	470	(a) 527	590	603	583	
	350	440	413	483	507	550			
	387	407	423	433	533				
	383	363	440	550	490				
		370	393	467	547				
		373	430	470	470				
		343	390	483	497				
		347	450	430	523				
		400	420	463	527				
		480	510	503	500				
		370	483	513	543				
		353	413	493	487				
		357	477	463	483				
		400	450	467	490				
		383	433	450	550				
		370	460	503					
		410	410	487					
		393	513	453					
		367	410	463					
			443	483					
			480	507					
			493	470					
			463	420					
			447	483					
			433	550					
			423	470					
			410	493					
			407						
			473						
Total	1540	7296	12793	12920	7674	1140	603	583	
Average	385	384	441.1	478.5	511.6	570	603	583	

Table 4

Average Pre- and Posttest Scores

Pretest Range	251-300	301-350	351-400	401-450	451-500	501-550	551-600	601-650
Pretest Mean	293.5	327.8	379.5	426.2	469.5	523.5	557	613
Posttest Mean	385	384	441.1	478.5	511.6	570	603	583
Raw Gain	91.5	56.2	61.6	52.3	42.1	46.5	46	-30

Even discounting the highest two "groups," which consist of one student each, it is apparent that students with pretest scores below 400 tend to gain more than 56 points, while those with pretest scores above 400 tend to gain less than 56 points. We can see the relationship of pretest to posttest more clearly by smoothing to a straight line with a kind of shortcut graph called a stem-and-leaf plot. In this way of laying out numbers, we arrange the first digits in order from the top and enter the final digit in its appropriate place to represent each number. For example, the scores 550, 543, 547, 533, 523, 527, 527 are represented as

55 0

54 3,7

53 3

52 3,7,7.

Note that since 527 appears twice, the 52 row has two 7's, one for each occurrence of 527.

This makes it clear at a glance that the median (middle value) of this set of data is 533 (the fourth score from the top or bottom) and that row 52 is the modal row (contains more observations than any rows in the above set of data).

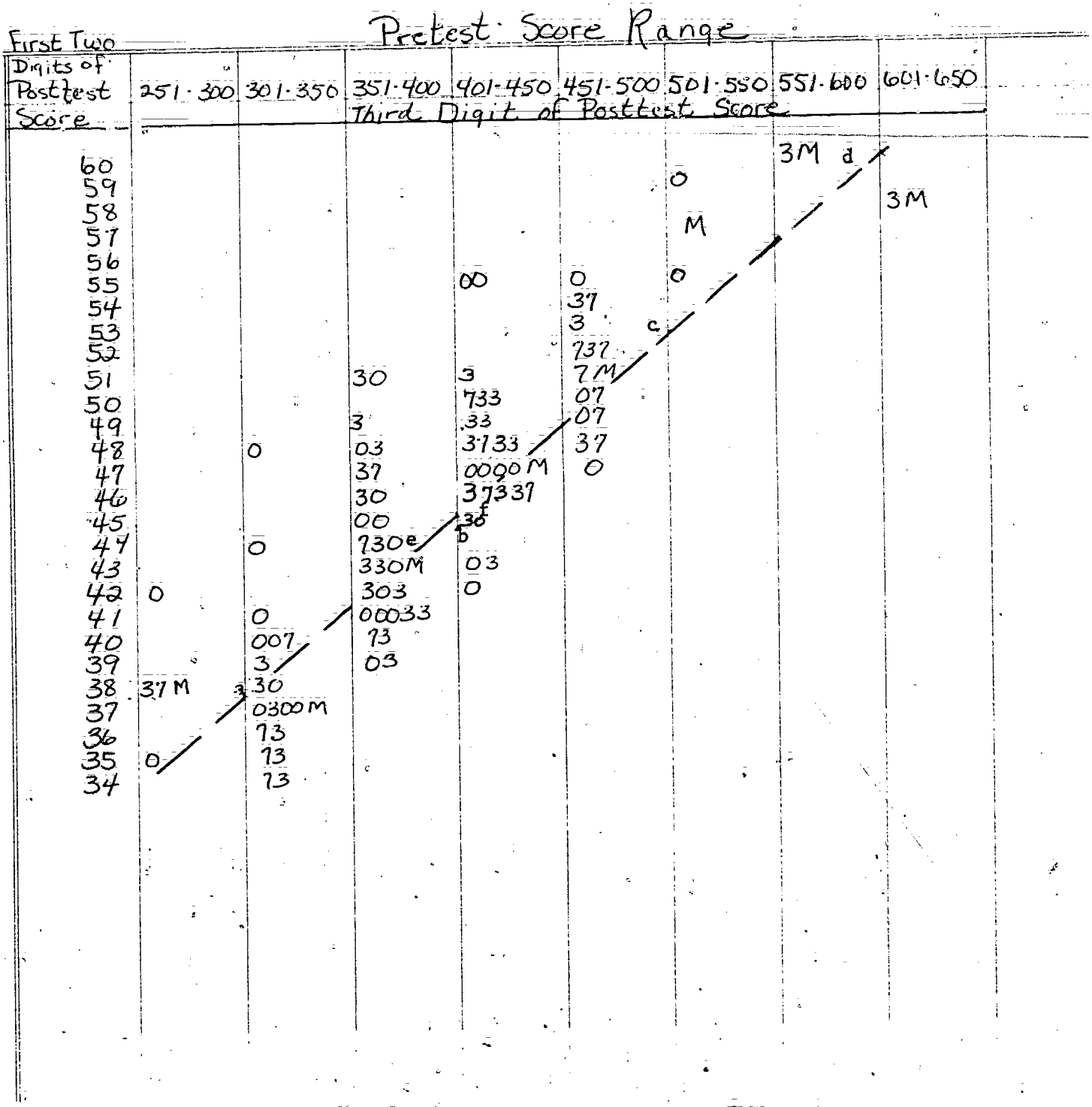
Table 5 shows the data of Table 3 recast as stem-and-leaf plots. In the first column, which represents the pretest score range 251-300, the four scores 420, 350, 387, 383 of Table 3 are represented as 0, 0, 3, and 7 in posttest rows 42, 35, and 38. The median of each column is designated "M." Although the M's do not fall exactly on a straight line, a prediction line that comes close to all of them has been drawn. At the 300/301 pretest b undary (a) this line is opposite a posttest score of 380, predicting an 80-point gain. The line cuts the 400/401 pretest

Table 5

Stem-and-Leaf Plots

Posttest Scores for Various Pretest Ranges

(Prototype)



boundary (b) at about 455, consistent with the 56-point observed mean gain. It intersects the 500/501 pretest boundary (c) at 530, predicting an average gain of 30 points for students with such high pretest scores. If we follow the line up to a pretest score of 600, (d), it predicts only about a five-point raw gain. Thus, if we use this rough prediction line, although the mean gain at a pretest score of 400 is about 55 points, each additional 100 points on the pretest is thus associated with about 25 points less raw gain, or 75 additional posttest points for each 100 pretest points. This line is said to have a "slope" of .75. The rough graph in Table 5 shows that for a group of students with pretest TOEFL scores of about 375, (e), the predicted posttest after one semester of intensive instruction was 430. However, almost one-fourth of this group scored above 470, and about one-fourth scored 410 or less. Average predictions can be guides to planning, but are far from predestination.

Predicting Scores for Subgroups

Based on these raw gain estimates, we can predict likely posttest scores for students with similar pretest scores in similar programs. With larger samples, we could begin to develop separate predictions for students with specific native language backgrounds and varying prior academic experience. It is likely that different prediction equations would fit students with comparable pretest scores from Indo-European vs. non-Indo-European backgrounds (particularly for TOEFL subscores), and students with extensive formal English study in their home country vs. those with less prior English study.

To illustrate comparison of two groups, we make use of the fact that the present sample of 98 students contained 68 first-semester students and 30 returning students, whose "pretest" scores had been obtained as posttests at the end of the summer, about two weeks before the beginning of the fall term. These 30 students are identified in the data of Appendix A by a 1 in column 79 of their first data card. One question that can thus be addressed by these data is whether continuing students are comparable to entering students in expected language growth. The means and standard deviations for the subgroup of continuing students are given in Table 6.

Table 6

Pre- and Posttest Means, Returning Students

	Mean	Standard Deviation
Pretest	409.93	48.04
Posttest	458.50	48.90

The pretest mean of this subgroup is about ten points above the overall pretest mean, and the posttest mean about three points above the overall posttest mean. The standard deviations are about 20 percent smaller than for the total group. Referring to the prediction line in Table 5 we see that these mean scores lie very close to the overall group prediction line. If we move about one-fifth of the way from 400 to 450, (f), corresponding to a pretest mean of 410, we find that the prediction line is about even with a posttest score of 460. Thus, the returning students' average scores fall only slightly below the prediction line for all students, and, from this rough analysis, it seems reasonable to assume that about the same relationship between pretest and posttest applies across groups. In Appendix B we will show how to develop a more precise prediction line, using a commonly available computer package (SPSS). A result of that analysis is explicit equations for the prediction lines. In Appendix B we will see that the equation for the overall line is

Posttest = 119.103 + .8411(pretest).

That is, the predicted posttest score corresponding to a given pretest score is .8411 times the given pretest score plus a constant, 119.103. For a pretest score of 200, the predicted posttest score is 287.323 (see Glossary, page 27).

For the returning students only (Appendix C), the equation is

Posttest = 113.32 + .8413(pretest),

a line that is parallel to the overall line, and about 5.8 points lower for a given pretest score. That more precise analysis suggests the possibility of a "diminishing return" of about six points for continuing students, but the small sample size (30 students) makes this no more than a possibility, which would have to be confirmed over several semesters in a particular program before it became a conclusion.

On the basis of the graphical analysis of this section, we would predict that a student entering this program with a TOEFL total of 300 would gain about 80 points in one semester, to reach a score of 380. Assuming that returning students' growth is similar to that of entering students, we would expect a gain of 60 points in the second semester, to reach a score of 440. In the third semester, we would expect a further gain of 45 points, to reach the score of 485. These predictions would be considerably improved by actually following returning students over several semesters, and by pooling several groups in similar curricula to obtain a larger sample. They could be somewhat improved by following the more precise estimation methods of Appendix B.

How much does this lower raw score gain for students with pretest scores above the mean have to do with instructional effectiveness, and how much with fallible measurement? We will address this problem in the following section.

Change and Regression to the Mean: An Example of the Problem

Lucy bowls in the Alley Oops League. The league average is 150. Last month she bowled a game of 100. If we know nothing else about Lucy, it is reasonable to assume that to some extent she was last month (a) a below average bowler, and (b) unlucky.

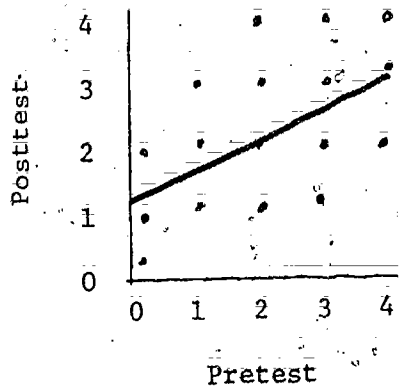
This month Lucy scored 200. In the absence of any further knowledge about Lucy it would be reasonable to assume that to some extent she is (a) an above average bowler, and (b) lucky this month. To what extent are these differences the result of changes in Lucy's ability, and to what extent are they due to luck? The question is important if we wish to compare the two scores. If chance differences of this size are common, we should conclude that she probably was and remains an average (150) bowler, and that there is no real difference in the two scores, beyond chance variation. On the other hand, if bowlers are usually quite consistent, seldom varying by more than 20 pins from one month to the next, we should conclude that there has probably been a real improvement in her bowling ability. Problems in measuring change center on the issue of assigning proportions of observed change to real growth and to luck. In Lucy's case, measured change is somewhere between the observed change of 100, if bowling scores are perfectly reliable, and 0, if bowling scores have no reliability. Reliability is just an indicator of the proportion of variance attributed to two possible sources of change. If half the change was due to ability growth, and half due to luck, we say the reliability is .5. If real ability growth represents a greater proportion of observed change, reliability is proportionally greater.

Test scores, like bowling scores, have ability and chance components. We can think of a "true" score analogous to a bowling average, and ask if it changes over time. If we wish to measure ability, we would prefer that luck played a small role in these considerations. But we recognize that the particular selection of questions on the test form, how well the student slept and ate the previous night and morning, and perhaps even the humidity or pollen count, add random errors to the observed score. There is a significant chance component in any test performance. This means that very low-scoring individuals are probably not really quite as low as they appear, because part of their low score was likely due to chance, or bad luck ("negative error," in measurement terms), while very high-scoring individuals are probably not really quite as high as they appear, because part of their high score was likely due to good luck ("positive error"). Even if no real change in anyone's true score takes place, initial low scorers will tend to score higher than their original scores on the posttest, and initial high scorers will tend to score lower on the posttest simply because random chance factors over the two times of measurement will tend to cancel out, making it less likely that the same person will be equally lucky twice in a row.

Figure 2 shows the predicted pattern when no real gain takes place (the posttest mean, like the pretest mean, is 2) but, as is always the case, measurement contains errors.

Figure 2

Changes from Pretest to Posttest (r = .5)



The range of the hypothetical pretest in Figure 2 is 0-4, but of the predicted posttest, only 1-3. The "missing" variance in the posttest is chance variance, which cannot be predicted. The observed scores on the posttest will probably still range from 0 to 4, but some who initially scored 1 will score 0 on the posttest, most will score 1 or 2, and some 3 or 4, giving an average of 1.5 instead of 1, while some who initially score 3 will score 0 or 1 on the posttest, most will score 2 or 3, and some 4, giving an average of 2.5 instead of 3. All posttest scores will shrink toward the average posttest score of 2.

Given this situation, the uncritical observer may ignore the chance variance and look at raw gains. If chance variance were 0, this would be appropriate. However, in the presence of error, raw gain is misleading. According to Figure 3, those with initial scores of 4 "lose" 1. This interpretation is wrong unless the test is perfectly reliable on both occasions. If reliability is less than perfect, such shrinking or "regression" toward the mean posttest score must happen on the average, even when no real change occurs.

One Solution to the Problem

How do we estimate true change to determine if programs are having uniform effects for students of differing entering abilities, or to determine the relative efficacy of different programs at different ability levels?

One approach is to estimate the reliability of the test in the group being studied, use this to predict the expected final score for each pretest score under the assumption that no real change has taken place, and call only observed discrepancies from this predicted score "change." Applying this approach to the situation in Figure 2, we would say that a student with an initial score of 0 and a final score of 1 had shown no real growth, since he had merely kept up with the pre-post difference to

be expected for a below-average initial score on a test of this reliability. On the other hand, a student with an initial score of 4 and a final score of 4 would be credited with a 1-point gain, since she had held her own against a predicted loss or "regression" of 1 point for this above-average pretest score. To illustrate the amount of change to be attributed to real gain rather than to regression, consider Lucy and her bowling scores once more.

If Lucy had bowled a 200 game immediately following her 100 game, with no time for intervening practice, we would be more inclined to conclude that her scores were highly variable, like those of the student who moved from 0 to 1 because of measurement error, and less inclined to attribute her 200 score in the following month to real improvement. On the other hand, if she had immediately followed her score of 100 with scores of 105 and 95, we would be more inclined to see her scores as quite consistent estimates of her true ability, with only a small component of chance variation, and be thus more likely to view a score of 200 one month later as reflecting true gain, with a similarly small component of chance variation.

Applying the Solution to Test Scores

We can apply this approach directly to the problem of measuring change in test scores. By following the administration of the pretest almost immediately with another test that we shall call a "reliability" test--given so soon (within about a week) after the pretest that little true change has had time to occur--we can estimate the change in observed scores likely to take place simply because of measurement error and test familiarization effects. For each pretest score, the average estimated score, or "regression line," on the reliability test forms a "no-change" baseline. Changes measured from this baseline, rather than "raw gain" measured from pretest scores themselves, give a more realistic picture of true gain, by taking measurement error and practice effects into account.

Imagine a group of students with a TOEFL pretest average score of 400. Suppose that at the end of a semester of intensive ESL instruction, the group average climbs to 450. If we looked at the group that scored around 300 at the pretest, we might find that their posttest mean was 370. The students who scored around 500 at pretest, on the other hand, might have moved to an average of 530. It would appear that those with a low pretest gain 70 points and those with a high pretest gain only 30 points.

The straight line regression equation (Figure 3) that fits these results is posttest = 130 plus .8(pretest).

That is, $370 = 130 + .8(300)$

$450 = 130 + .8(400)$

$530 = 130 + .8(500)$.

Here 130 is the "intercept" or predicted value at the posttest for a hypothetical pretest score of 0, and .8 is the "slope." With a slope of .8, the predicted posttest score goes up eight points for every additional ten pretest points. In the presence of measurement error and parallel tests, it is unusual for the slope to be as great as 1.0, unless the variance of the posttest is much larger than the variance of the pretest. A slope greater than 1.0 can occur, however, if gains are positively correlated with pretest, that is, if those who start with higher scores really gain much more than do those with lower scores. In such cases, any regression to the mean would be counteracted by the large increase in posttest variance. If pretest and posttest variances are about the same, however, the slope will usually range from .65 to .95. In such more usual cases as this example, with slope = .8, even with equal growth across the score range, students with scores below 400 would be pushed up toward the mean by unreliability, and students with pretest scores above 400 would be pushed down toward the mean.

If a reliability test administered immediately after the pretest were to yield a parallel baseline (same slope as posttest), e.g.: test(B) = 90 + .8(pretest) (Figure 4), the no-change baseline for students with a pretest score of 300 would be 330 and for those with a pretest score of 500, 490. The estimated true gain from 330 to 370 and from 490 to 530 points, respectively, would then be a uniform 40 points at both points in the score range, even though raw gain appears greater for lower scores (Figure 3).

If the prediction equation for the no-change baseline had been found instead to have a flatter slope, such as .75, e.g.: test(B) = 100 + .75(pretest), the predicted baseline score for a pretest score of 300 would then become 100 + .75(300) = 325. For a pretest score of 500, the baseline would be 100 + .75(500) = 475. The posttest gain from the baseline for a pretest score of 300 would be 370 - 325 = 45, and for a pretest score of 500, 530 - 475 = 55. In this case, even though raw gain was less at higher scores, corrected gain would be slightly greater at higher scores (Figure 5).

Figure 3

Raw Gain - less at higher pretest scores

AB: Posttest = 130 + .8 x (pretest)
CD: Baseline = pretest

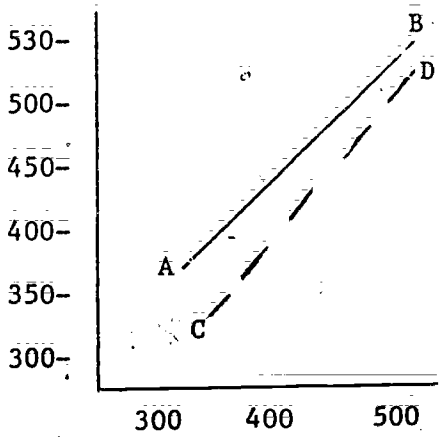


Figure 4

Gain from No-Change Baseline - uniform across pretest scores

AB: Posttest = 130 + .8 x (pretest)
EF: Baseline = reliability test prediction line = 90 + .8 x (pretest)

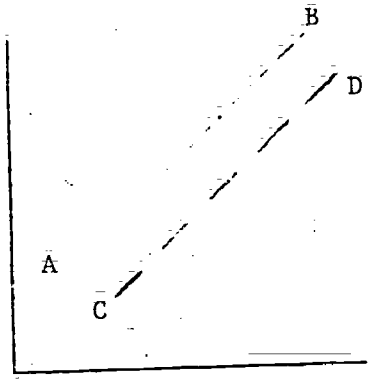
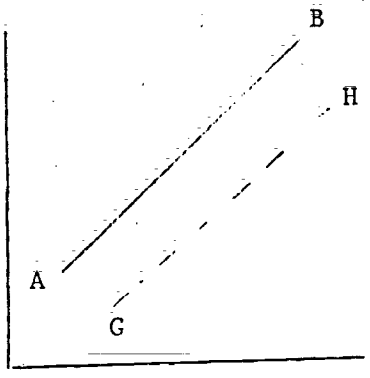


Figure 5

Gain from No-Change Baseline - greater at higher pretest scores

AB: Posttest = 130 + .8 x (pretest)
GH: Test(2) = 100 + .75 x (pretest)



We can perform these gain regression analyses graphically, by comparing posttest scores with baseline scores predicted on a plot like that of Table 5. However, with the wide availability of computer facilities, and of students and staff experienced in using statistical analysis packages, the use of a computer regression analysis program is suggested. In the following pages, we summarize the results of a computer analysis applying the above solution of estimating change from a reliability test baseline to the data from our sample of students. Details of instructing the computer to perform these analyses, and of the resulting output, are in Appendix B.

Regression Using a Computer Package

Appendix B gives the details of a regression analysis for the total TOEFL score and for each of the three TOEFL subscores: Listening Comprehension, Structure and Written Expression, and Reading Comprehension and Vocabulary. The Statistical Package for the Social Sciences is used for illustrative purposes, but other packages (e.g., SAS, BMD, or Data Text) or locally available linear regression programs would serve as well. Staff at your computer facility can tell you which program is most economical for your data and analysis needs.

The results of the sample regression analyses are summarized and briefly interpreted in this section.

Each of the analyses is based on three tests: pretest A, given at or before the beginning of the semester; reliability test B, given one week after the beginning of the semester, and posttest C, given at the end of the semester. The purpose of the reliability test was to establish a no-change baseline, making it possible to estimate the apparent "growth" to be expected from measurement error and test familiarization or practice effects. This baseline was established by relating test B to pretest A, to determine an average prediction line. At posttest C, student growth was assessed by comparing a student's test C score with his or her predicted test B score (as predicted from pretest A), rather than with the original pretest A score. Changes from A to predicted B scores were assumed to result from factors other than instruction, and were thus discounted in estimating gains due to instruction. Regression equations are based on means, rather than medians. Thus, the few high scores in Table 3 have more influence for those estimates than was the case in the graphical approximation.

For the total TOEFL scores, the reliability test B showed the following predicted scores for various pretest A scores.

Table 7

Total Pretest and Reliability Scores

Reliability Test B	323	371.5	419.9	468.4	516.8	565.3	613.7
Pretest A	300	350	400	450	500	550	600

Thus, with no real change, lowest-scoring students would appear to gain 23 points and highest-scoring students, only about 14 points.

This differential is in the direction discussed previously, but it is not a large discrepancy. Indeed, when we look at posttest scores (Table 8), we find the corrected gain on total scores remains larger for students with lower pretest scores.

Table 8

Total Scores and Gains

Posttest $\hat{C}$	371.4	413.5	455.5	497.6	539.6	581.7	623.7
Reliability Test $\hat{B}$	323	371.5	419.9	468.4	516.8	565.3	613.7
Pretest A	300	350	400	450	500	550	600
Raw Gain $\hat{C}$ - A	71.4	63.5	55.5	47.6	39.6	31.7	23.7
Corrected Gain $\hat{C}$ - $\hat{B}$	48.4	42	35.6	29.2	22.8	16.4	10.0

Even after correcting for test reliability, a student with a pretest score of 300 is estimated to gain over 50 more than a student with a pretest score of 400. The graphs of these relationships are given in Figures 6 and 7.

However, examination of the subtest scores reveals that most of this differential growth is concentrated in one TOEFL subtest, Listening Comprehension, subtest 1. Table 9 gives predicted C1 and B1 scores for various A1 scores:

Table 9

Listening Comprehension Scores and Gains

Posttest $\hat{C}$	41.7	44.9	48.2	51.4	54.7	57.9	61.2
Reliability Test $\hat{B}$	34.4	38.3	42.2	46.2	50.1	54.0	57.9
Pretest A	30	35	40	45	50	55	60
Raw Gain $\hat{C}$ - A	11.7	9.9	8.2	6.4	4.7	2.9	1.2
Corrected Gain $\hat{C}$ - $\hat{B}$	7.3	6.6	6.0	5.3	4.6	3.9	3.3

LANGUAGE GAIN ANALYSIS

08/10/81

FILE NONAME (CREATION DATE = 08/10/81)

SCATTERGRAM OF (DOWN) BTOT (ACROSS) ATOT

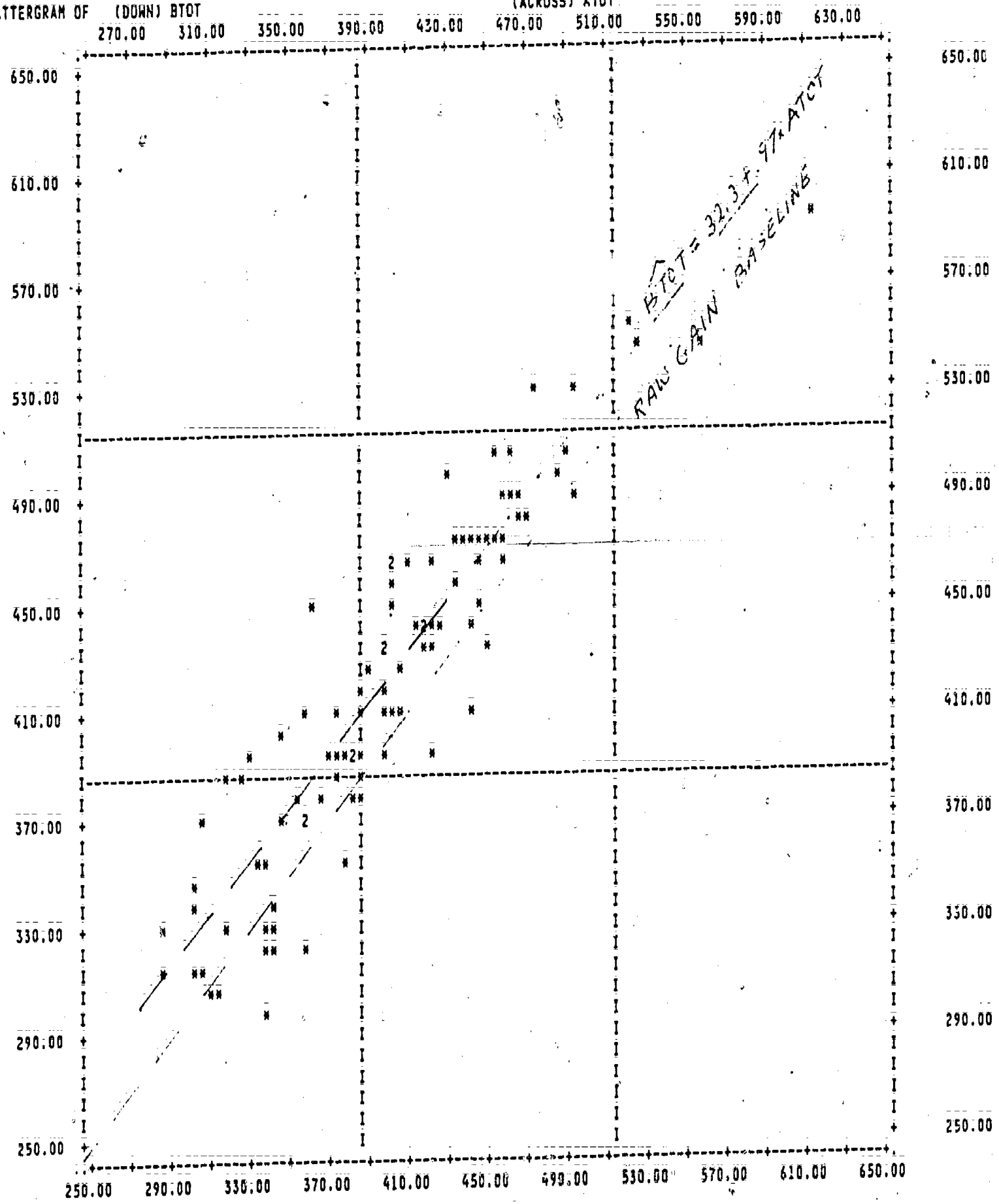


Figure 6

LANGUAGE GAIN ANALYSIS 08/10/81

FILE NONAME (CREATION DATE = 08/10/81)
SCATTERGRAM OF (DOWN) CTOT (ACROSS) ATOT

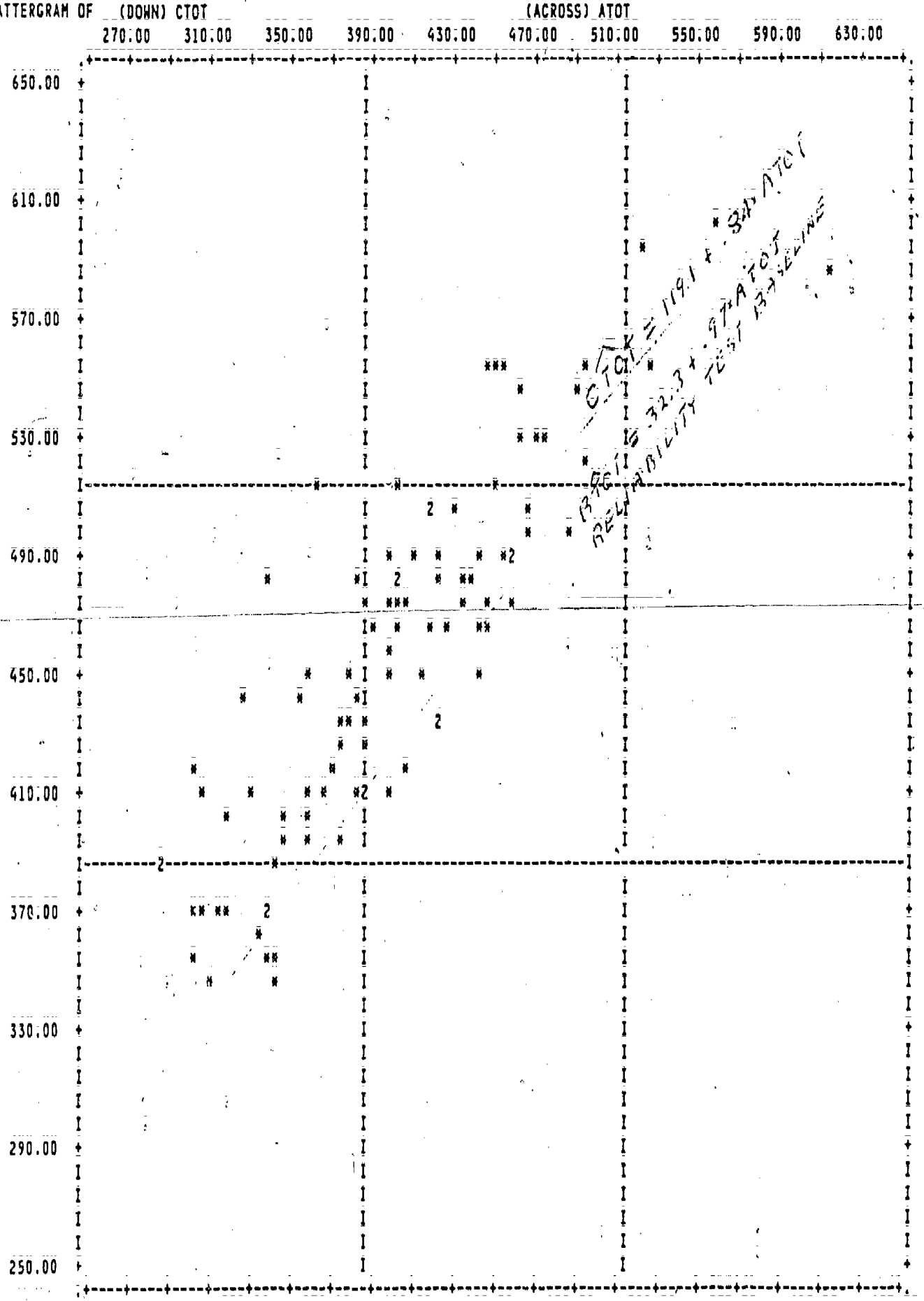


Figure 7

The corrected Listening Comprehension gains are thus seen to follow a pattern similar to that of the TOEFL total scores, with lower-proficiency students registering considerably larger gains than those achieved by students with higher pretest scores.

Table 10 gives the pattern for subtest 2, Structure and Written Expression.

Table 10

Structure and Written Expression Raw and Corrected Gains

Posttest $\hat{C}$	36.8	40.5	44.2	47.9	51.7	55.4	59.1
Reliability Test $\hat{B}$	34.2	38.0	41.8	45.7	49.5	53.3	57.1
Pretest A	30	35	40	45	50	55	60
Raw Gain $\hat{C} - A$	6.8	5.5	4.2	2.9	1.7	0.1	- .9
Corrected Gain $\hat{C} - \hat{B}$	2.4	2.5	2.4	2.2	2.2	2.1	2.0

Although the raw gain is dramatically less for students with high pretest scores, even becoming negative for a pretest score of 60, the corrected gain is seen to be nearly uniform across the score range of the Structure and Written Expression subtest. This subtest does not contribute materially to the lessened gain for students with high pretest noted for the total TOEFL scores.

Table 11 gives raw and corrected gains for subtest 3, Reading Comprehension and Vocabulary.

Table 11

Reading Comprehension and Vocabulary Raw and Corrected Gains

Posttest $\hat{C}$	35.9	40.4	44.8	49.3	53.7	58.2	62.6
Reliability $\hat{B}$	32.8	37.5	42.2	46.9	51.6	56.3	61.0
Pretest A	30	35	40	45	50	55	60
Raw Gain $\hat{C} - A$	5.9	5.4	4.8	4.3	3.7	3.2	2.6
Corrected Gain $\hat{C} - \hat{B}$	3.1	2.9	2.6	2.4	2.1	1.9	1.6

Although higher proficiency students gain slightly less on this subtest, corrected gains across the middle range of scores are again reasonably uniform. The Reading Comprehension and Vocabulary subtest thus does not contribute strongly to the diminishing growth found for higher-proficiency students with the TOEFL total scores. We can conclude that the observed pattern in the TOEFL total scores stems primarily from the lower growth for intermediate and higher level students observed in the Listening Comprehension subtest scores. It may be that new students' growth in comprehension of the English phonological system is quite rapid, but, after this is achieved, that further growth in Listening Comprehension depends on the same factors that influence Structure and Vocabulary.

This interpretation is supported by the Listening Comprehension results for the 30 returning students presented in Appendix C. Listening Comprehension posttest scores for this subgroup are given in Table 12.

Table 12

Listening Comprehension Raw and Corrected Gains
30 Continuing Students

Posttest $\hat{C}$	37.4	41.4	45.4	49.4	53.4	57.4	61.3
Reliability Test $\hat{B}^*$	34.4	38.3	42.2	46.1	50.1	54.0	57.9
Pretest A	30	35	40	45	50	55	60
Raw Gain $\hat{C}$ - A	7.4	6.4	5.4	4.4	3.4	2.4	1.3
Corrected Gain $\hat{C}$ - $\hat{B}$	3.0	3.1	3.2	3.3	3.3	3.4	3.4

*Based on total group data equation

Growth is almost perfectly uniform across the scale for this group. It appears that if, after a semester of familiarization with spoken English, a student who still has a low Listening Comprehension score is not likely to exhibit the rapid growth shown by newly entering students.

Summary

Because of measurement error, raw gain scores (i.e., simple differences of posttest and pretest scores) tend to overestimate real growth for initially low-scoring students and to underestimate gain for initially high-scoring students. By following a pretest with a reliability administration, it is possible to estimate the probable apparent change due to unreliability and practice effects, and to discount these in

estimating gain due to instruction. The technique requires that we fit a prediction line to the test B (reliability test) scores corresponding to each pretest score, and consider change due to instruction at posttest to be measured by deviations from this prediction line, rather than from pretest or reliability test observed score. For example, a given student might score 300 on the pretest (test A) and 330 on test B. If the regression equation relating all test B to test A scores were found to be test B = 100 + .8(test A), the student's expected score on test C, assuming no further change, would be the test B prediction: 100 + .8(300) = 340. If the actual test C score were 400, we would estimate gain as 400 - 340 = 60 points, rather than the raw test C - test A gain of 100.

References

Nie, N. H., Hull, C. H., Jenkins, J. G., Sternbrenner, K., & Bent, D. H. SPSS: Statistical Package for the Social Sciences (2nd ed.). New York: McGraw Hill, 1970.

Glossary

Analysis of covariance--the method used to determine the shared variation of two or more related variables. For example, a pretest is used as a covariate to predict a posttest. Then standard analysis of variance is used to estimate the effect of a treatment on the residual variation in posttest scores, not accounted for by the pretest.

Average--the sum of the measures, items, scores, etc., divided by their number or frequency.

Chance variation--the variation that one would expect from the scores of equivalent forms given close in time without instruction or mistakes.

Correlation--the amount of similarity in degree and direction between two sets or ranks of variables; a measure of the degree to which knowledge of one set allows us to predict the other set.

Equivalent forms--two or more forms of a test that are so similar they can be used interchangeably and yet are not identical; two or more test forms that yield about the same mean and variability of scores, and whose items are similar with respect to type, difficulty, distribution of item-test correlations, and representative coverage of content.

Mean (average)--the sum of the measures, items, scores, etc., divided by their number or frequency.

Measurement error (standard error)--the deviation from the true score that is due to chance variation. For a given observed score, the specific value of the measurement error is unknown, but the average error of a set of scores describes their precision.

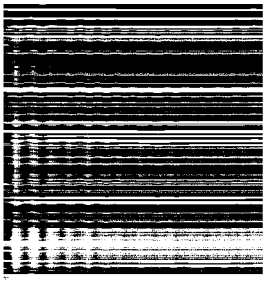
Median--the middle score in a distribution or set of ranked scores; the point (score) that divides the group into two equal parts; the 50th percentile; a measure of central tendency.

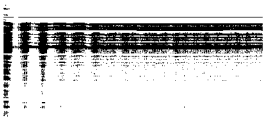
Mode--the score or value that occurs most frequently in a distribution; a measure of central tendency.

Posttest--a test given at the conclusion of an educational project or treatment to determine posttreatment status of the examinee or group in regard to some skill, aptitude, or achievement.

Pretest--a test given to determine the status of the examinee or group in regard to some skill, aptitude, or achievement, as a basis for judging the effectiveness of subsequent treatment.

Probability--if there is a known number, p, of possible occurences of an event and q possible nonoccurences, and if each of the total, p + q, possible outcomes is equally likely, then the probability of the event is $\frac{p}{p+q}$.





Regression effect or regression to the mean--tendency of a predicted score to be nearer to the mean of its distribution than the score from which it is predicted is to its mean. Because of the effects of regression, students making extremely high or extremely low scores on a test tend to make less extreme scores, i.e., closer to the mean, on a second administration of the same test or on some predicted measure. In general, the greater the errors of measurement and prediction, the more pronounced is the regression effect. For example, the heights of parents and of their children are related, but one cannot be perfectly predicted from the other. If we select the ten tallest individuals in the world, it is extremely likely that their average height exceeds the average height of their parents, but it is also extremely likely that their average height will exceed the average mature height of their children. This will be true even if the average height of the entire population is increasing slightly from one generation to the next.

Regression (line)--if two paired lists of numbers, say pretest scores and posttest scores, are plotted in two dimensions, say pretest horizontally and posttest vertically, then there is exactly one straight line that can be drawn through the plot so that it passes closest to the means of all those sets of posttest scores that correspond to each pretest score. On the average, for all pretest scores, this is the best straight-line fit to the observed posttest scores.

Reliability--the extent to which a test is consistent in measuring whatever it does measure; dependability, stability, trustworthiness, relative freedom from errors of measurement.

Slope--the steepness of ascent of a straight-line graph. If the line is described by the equation $Y = mX + c$, where Y represents the vertical axis, the value of Y will increase m units for each unit increase in X, and m is the slope. For example, if $Y = 0.5X + 100$, and X increases from 200 to 300, Y will increase half as much, from 200 to 250.

Standard deviation--a measure of the variability of dispersion of a distribution of scores. The more the scores cluster around the mean, the smaller the standard deviation. For a normal distribution, about two-thirds (68.3 percent) of the scores are within the range from one S.D. below the mean to one S.D. above the mean.

True score--a score entirely free of error--hence, a hypothetical value that can never be obtained by testing, which always involves some measurement error. A "true" score may be thought of as the average score from an infinite number of measurements from the same or exactly equivalent tests, assuming no practice effect or change in the individual during the testings.

Variance--a measure of variability equal to the square of the standard deviation; the average of the squared deviations from the mean. The variance of the sum of independent random variables is the sum of their variances. This makes the measure useful in theory. For practical purposes, the percent of the standard deviation explained may be more meaningful than is the percent of variance explained.

Appendix A

Setting Up the Data

Three scores will be available for each test being analyzed (we will discuss missing data later). If these scores have been entered in a cumulative record form, such as that illustrated in Table 1 of the text, a single card can be punched for each student. This card will contain the student ID number, program code, and pretest, reliability test, and posttest scores in a fixed set of columns. Additional information, such as native language, number of years of English study, or teacher ratings may also be included for further analyses.

If the scores have not been copied onto a common record form, but are on separate lists, it is often easier to punch up one card for each testing occasion, with student ID number, score for a given test in the same column on each card, and a test occasion number (1 = pretest, 2 = reliability test, 3 = posttest) in column 80. Additional variables need be punched only on card one. The three resulting decks of cards may then be stacked in order with deck one on top, and run through a card sorter once for each column of the student ID number, starting with the rightmost identification digit. The resulting merged deck will have cards in order within student ID's in numerical order. Listing this deck makes it easy to spot breaks in the 1, 2, 3 sequence and to pull out cards for students who missed one or more testings. In our example, student ID's (numbers ranging from 001 to 111) are punched in the first three columns of each card. Column 4 is left blank, and the three TOEFL subscores and TOEFL total are punched sequentially on a separate card for each of the three tests. Listening subscores are in columns 5 and 6, Structure and Written Expression in columns 7 and 8, Reading Comprehension in 9 and 10, and TOEFL total scores in 11 through 13. Column 80 contains the testing occasion, and, because some students were pretested on an earlier date than others, column 79 identifies such students with a numeral 1. The cards were sorted and incomplete sets removed. This resulted in complete data for 98 students. A listing of the control and data cards is given in Table A-1. The explanation of the control cards, such as variable list input format and scattergram, is given in Appendix B.

Table A-1

```
RUN NAME        LANGUAGE GAIN ANALYSIS
VARIABLE LIST   A1,A2,A3,ATOT,B1,B2,B3,BTOT,C1,C2,C3,CTOT
INPUT MEDIUM    CARD
INPUT FORMAT    FIXED(4X,3F2.0,1F3.0/4X,3F2.0,1F3.0/4X,3F2.0,1F3.0)
N OF CASES      98
PEARSON CORR    A1 TO CTOT
OPTIONS         5
STATISTICS      1
READ INPUT DATA
001 544444473                                                        1
001 575447527                                                        2
001 555053527                                                        3
002 423431357                                                        1
002 464137413                                                        2
002 505536403                                                        3
003 433232357                                                        1
003 423831370                                                        2
003 504133413                                                        3
004 484646467                                                        1
004 484947480                                                        2
004 525248507                                                        3
006 513733403                                                        1
006 524542463                                                        2
006 584043470                                                        3
050 512734373                                                        1
050 463237383                                                        2
050 503641423                                                        3
007 473433380                                                       11
007 423740397                                                        2
007 543840440                                                        3
008 433633373                                                       11
008 394038390                                                        2
008 414037393                                                        3
009 393032337                                                       11
009 352629300                                                        2
009 443235370                                                        3
010 462428327                                                        1
010 493334387                                                        2
010 573639440                                                        3
011 362531307                                                        1
011 433334367                                                        2
011 483638407                                                        3
013 483736403                                                        1
013 524146463                                                        2
013 594244483                                                        3
014 382834333                                                       11
014 383930357                                                        2
014 423037363                                                        3
015 352733317                                                        1
015 393228330                                                        2
015 423732370                                                        3
017 502537373                                                       11
017 464136410                                                        2
017 534036430                                                        3
018 504036420                                                        1
018 443539393                                                        2
018 504139433                                                        3
019 413531357                                                       11
019 393027320                                                        2
019 444033390                                                        3
```

020	362727300	1
020	393329337	2
020	513837420	3
021	453737397	1
021	494135417	2
021	524340450	3
022	312833307	1
022	353228317	2
022	443533373	3
023	443433370	11
023	453934393	2
023	533637420	3
024	525252520	1
024	595355557	2
024	645657590	3
025	372033300	1
025	393232343	2
025	403035350	3
026	433839400	1
026	453544413	2
026	564156510	3
027	484541447	1
027	524248473	2
027	625251550	3
028	483842427	11
028	484540443	2
028	544244467	3
029	363729340	1
029	393032337	2
029	353236343	3
030	413935383	1
030	432743377	2
030	544348483	3
033	383025310	1
033	342631303	2
033	393530347	3
034	533942447	11
034	514445467	2
034	544542470	3
036	303134317	1
036	354437387	2
036	423840400	3
037	453536387	11
037	423536377	2
037	443842413	3
038	413030337	11
038	393327330	2
038	453633480	3
039	434345437	1
039	543752477	2
039	603748483	3
040	434241420	1
040	414446437	2
040	414444430	3
042	504439443	11
042	544435443	2
042	524542463	3
043	504643463	11
043	514951503	2
043	525850533	3

044	444338417	1
044	504241443	2
044	584944503	3
045	595553557	1
045	595154547	2
045	645859603	3
046	332231287	1
046	392536333	2
046	383345387	3
048	514145457	11
048	514147463	2
048	534945490	3
051	606361613	1
051	616157597	2
051	605956583	3
052	322831303	1
052	362631310	2
052	423633370	3
041	483948450	11
041	493347430	2
041	594253513	3
053	545142490	1
053	555244503	2
053	585353547	3
054	392735337	1
054	383533353	2
054	423331353	3
055	443636387	1
055	494134413	2
055	564542477	3
056	534139443	11
056	504447470	2
056	574447493	3
057	433529357	11
057	374034370	2
057	504540450	3
058	504242447	1
058	494441447	2
058	544144463	3
059	483839417	1
059	464244440	2
059	504545467	3
060	513934413	11
060	544137440	2
060	524538450	3
061	564239457	1
061	594543490	2
061	574440470	3
062	383332343	1
062	393029327	2
062	413531357	3
063	353930347	1
063	393735370	2
063	453738400	3
064	553440430	1
064	583952497	2
064	584647503	3
065	483830387	1
065	474236417	2
065	523840433	3

066	433838397	1
066	453543410	2
066	533748460	3
067	414035387	1
067	384037383	2
067	483837410	3
068	274240363	1
068	524141447	2
068	595144513	3
069	544343467	1
069	534449487	2
069	594347497	3
070	433833380	1
070	413838390	2
070	414339410	3
073	333831340	1
073	402729320	2
073	413638383	3
074	453741410	11
074	524741467	2
074	505145487	3
075	484539440	1
075	463937407	2
075	534241453	3
076	544549493	1
076	544250487	2
076	555250523	3
077	584555527	11
077	594757543	2
077	594759550	3
078	393235353	1
078	423437377	2
078	504340443	3
080	485043470	1
080	455544480	2
080	565349527	3
082	424039403	1
082	474443447	2
082	524047463	3
083	494851493	1
083	505455530	2
083	555260557	3
084	534152487	11
084	574448497	2
084	544353500	3
085	514345463	11
085	544646487	2
085	595252543	3
086	312233287	1
086	352731310	2
086	443536383	3
087	343228313	1
087	362531307	2
087	423633370	3
088	484138423	11
088	524536443	2
088	574345483	3
089	443640400	1
089	454844457	2
089	544149480	3

090	564139453	11
090	544641470	2
090	574643487	3
091	444332397	1
091	464241430	2
091	564943493	3
092	383031330	1
092	424035390	2
092	444138410	3
093	444239417	11
093	434245433	2
093	564749507	3
095	424233390	1
095	444539427	2
095	485041463	3
096	453236377	11
096	343835357	2
096	494540447	3
097	433634377	1
097	423937393	2
097	564133433	3
098	483836407	1
098	493841427	2
098	573846470	3
099	483830387	11
099	463437390	2
099	474337423	3
100	503735407	11
100	404537407	2
100	523539420	3
101	363830347	1
101	414435400	2
101	463537393	3
102	504337433	1
102	544543473	2
102	544942483	3
103	383537367	11
103	413339377	2
103	433644410	3
104	474345450	1
104	494548473	2
104	545655550	3
105	464843457	1
105	495142473	2
105	544944490	3
106	483239397	1
106	443736390	2
106	533534407	3
107	434036397	1
107	464639437	2
107	554641473	3
108	484642453	1
108	544850507	2
108	635448550	3
109	533839433	11
109	524541460	2
109	544443470	3
110	393131337	1
110	353130320	2
110	463331367	3

```
111 473940420                                                                    1
111 494744467                                                                    2
111 544747493                                                                    3
SCATTERGRAM    BTOT(250,650),CTOT(250,650) WITH ATOT(250,650)
STATISTICS     ALL
SCATTERGRAM    B1(20,70),C1(20,70) WITH A1(20,70)
STATISTICS     ALL
SCATTERGRAM    B2(20,70),C2(20,70) WITH A2(20,70)
STATISTICS     ALL
SCATTERGRAM    B3(20,70),C3(20,70) WITH A3(20,70)
STATISTICS     ALL
FINISH
```

Appendix B

Analyzing and Interpreting the Data

Reading the Data

To enable the computer to read the information, the names of the variables and where to find them must be given. In SPSS, this is accomplished with four cards, immediately following the required RUN NAME card.

The first, VARIABLE LIST, gives a list of variable names separated by commas. We have four scores on each of three cards, and have decided to call them A1, A2, A3, ATOT, B1, B2, etc. These labels are punched on the VARIABLE LIST card beginning in column 16.

The second card, INPUT MEDIUM CARD, with "CARD" beginning in column 16, is self-explanatory.

The third card, INPUT FORMAT, indicates the location of each variable in order on the cards. Beginning in column 16, with FIXED, the card contains a FORTRAN format statement (4X, 3F2.0, 1F3.0,/4X, 3F2.0, 1F3.0/4X, 3F2.0, 1F3.0). This code instructs the computer to skip the first four ID spaces, read three two-digit numbers (which it will assign to variables A1, A2, and A3), read a three-digit number, which it will assign to variable ATOT, skip to the next card, read the four B variables from the reliability test in the same format, skip to the third card, and read the posttest C scores to complete one case.

The final control card in this sequence, N OF CASES 98, indicates the number of times this procedure must be repeated to complete reading the data.

Performing the Analyses

The basic descriptive statistics--means, standard deviations, and correlations--are obtained with a single set of control cards. The equations of the prediction lines are obtained with a set of control cards for each test score.

The basic descriptive statistics--means, standard deviations, and correlations--are obtained with a single set of control cards:

```
PEARSON CORR    A1    TO CTOT

OPTIONS         5

STATISTICS      1
```

These are followed by the

READ INPUT DATA card, and by the data deck.

The first page of the computer output, listing the control cards and showing which card columns were read for each variable, is reproduced as Table B-1 (page B.10).

The resulting univariate statistics and correlations are given in Tables B-2 to B-4.

The release of SPSS used at the ETS computer facility requires that subsequent analysis request cards after the first set follow the data deck. This requirement may vary with other releases of the SPSS package.

The next analysis performed (total test score) is a prediction of the total reliability test, B, from the pretest scores, A, to establish a no-change baseline. This is followed by a prediction of total posttest scores, C, from A.

Change for each pretest score is the difference between C and B: predicted posttest score minus predicted baseline for that value of the pretest. Both analyses may be perfcrmed with a two-card request, immediately following the data deck:

```
SCATTERGRAM     BTOT (250,650), CTOT (250,650) WITH ATOT (250,650)
STATISTICS      ALL                                    (Table B-5)
```

This first shows the relationship of BTOT with ATOT, followed by the estimates needed for the baseline prediction equation. Then the relationship of CTOT to ATOT is plotted, followed by the estimates required for the posttest prediction equation. The ranges (250,650) scale the plots for easier readability. The range (200,700) would also work.

Table B-6 gives the plot of the BTOT observations (on the vertical axis) corresponding to each observed value of pretest score for each of the 98 students. Each asterisk represents one student's pair of A and B total scores. The numeral 2 represents pairs of scores occurring for two different individuals.

The scatterplot shows a strong, quite linear relationship between pretest and reliability test, with a concentration of scores in the lower two-thirds of the two score ranges. The 297-point range of the reliability test, 300-597, is slightly less than the 326-point range of the pretest.

Table B-7 gives the information necessary to determine the prediction equation for the no-change baseline. The two underlined quantities, INTERCEPT (A) = 32.33932 and SLOPE (B) = 0.96865, give the constants of the baseline equation: BTOT = 32.339 + .969ATOT.

This line has been drawn on the scatterplots of Tables B-6 and B-8.

Plotting the No-Change Baseline

The baseline can be plotted by choosing any two convenient values of ATOT--300 and 600, for example--calculating the corresponding predicted values of BTOT from the prediction equation, and plotting the two ATOT, BTOT values on the graph. Thus,

$$\widehat{BTOT}\ (1) = 32.339 + .969 \times 300 = 323.0, \quad \text{and}$$

$$\widehat{BTOT}\ (2) = 32.339 + .969 \times 600 = 613.7$$

As a check, it is a good idea to choose another value, say ATOT (3) = 500, yielding BTOT (3) = 516.8. The three points should lie on a single straight line. If they do not, recheck the calculations and plotting.

Table B-8 gives the scatterplot relating pretest and posttest scores. The range (343-603) of posttest scores has further diminished to 260 points. The lowest posttest score, 34, is 56 points higher than the lowest pretest score, and the highest posttest score is 10 points lower than the highest pretest score.

Table B-9 gives the estimates for the constants in the equation predicting posttest from pretest:

$$\widehat{CTOT} = 119.103 + .841ATOT$$

This equation predicts that a student with a pretest score, ATOT, of 300 would be expected to score around 371.4 at posttest, CTOT, a raw gain of 71.4 points, while a student with a pretest score of 500 would be expected to score around 539.6, a raw gain of 39.6, or about 32 points less than that anticipated for the lowest-scoring students.

If we graph this prediction equation and compare it with the slope = 1 "posttest = pretest" dashed baseline that is implicit in using raw gain, it shows that initially low-scoring students gain much more than do initially high scorers. Instead of comparing raw gains, however, we wish to compare gains from baseline. Comparing posttest scores with the no-change baseline BTOT = 32.339 + .969 x 300 yields an estimated gain for a student with a pretest score of 300 of CTOT - BTOT = 371.4 - 323.0 = 48.4 points. For a student with an initial score of 500, the estimated gain is 539.6 - 516.8 = 22.8 points. Thus, although the discrepancy is reduced from the raw gain difference of 32 to a corrected difference of 26 points, it appears that regression effects are not sufficient to account for the discrepancy in gain across the score range for this particular sample of students.

Analyzing Subtests

In the example, the request for analysis of total scores was followed by the analysis request for the Listening Comprehension subtest:

SCATTERGRAM B1 (20,70), C1 (20,70) WITH A1 (20,70)*

STATISTICS ALL (Table B-10)

The resulting plots and statistics are given in Tables B-11 to B-14. Again, the plots are quite linear, with a large displacement from the B1 vs. A1 plot to the C1 vs. A1 plot, suggesting considerable real growth over the time interval from B1 to C1.

The equation of the baseline is

$\hat{B1} = 10.959 + .782A1$ (Table B-12. Find it! It's not underlined this time.)

For the posttest, the prediction equation is

$\hat{C1} = 22.167 + .650A1$ (Table B-14)

A student with a pretest score of 30 would thus be expected to achieve a B1 score of

$10.96 + .782 \times 30 = 34.42$, and a C1 score of

$22.17 + .650 \times 30 = \underline{41.67}$ for an estimated gain of
7.25

A student with a pretest score of 60 would be expected to achieve a B1 score of

$10.96 + .782 \times 50 = 50.06$ and a C1 score of

$22.17 + .650 \times 50 = \underline{54.67}$ for an estimated gain of
4.61

These lines have been drawn on the scatterplots (B-11 and B-13) and contrasted with the dashed raw-gain baseline. With Listening Comprehension, as with TOEFL total, it appears that the initially low-scoring students did in fact gain more in this class than did those who started with higher scores, even after measurement errors and practice effects are taken into account. The estimated gain is positive across the scale, even though "raw gain" is negative for students with pretest scores above 50.

*The (-20,70) scaling is again to improve the readability of the graph. Without it, the A1 axis would be in units of 27.0, 30.3, 33.6, etc.

The next analysis request calls for information about the Structure and Written Expression subtest:

SCATTERGRAM B2 (20,70), C2 (20,70) WITH A2 (20,70)

STATISTICS ALL (Table B-15)

The resulting plots (Tables B-16 and B-18) show an unexpectedly tighter scatter of points for C2 vs. A2 than for B2 vs. A2, even though C2 and A2 are more greatly separated in time. The higher correlation, .803, for C2 and A2 vs. .760 for B2 and A2 confirms this visual impression. The baseline equation is

$\hat{B2} = 11.247 + .765 \times A2$ (Table B-17)

The equation predicting posttest from pretest is

$\hat{C2} = 14.541 + .742 \times A2$ (Table B-19)

A student with a Structure and Written Expression pretest of 30 would have an expected baseline score

$\hat{B2} = 11.247 + .765 \times 30 = 34.20$ and posttest

$\hat{C2} = 14.541 + .742 \times 30 = \underset{2.60}{36.80}$ for an estimated gain of

A student with a Structure and Written Expression pretest score of 50 would have an expected baseline of

$\hat{B2} = 11.247 + .765 \times 50 = 49.50$ and posttest

$\hat{C2} = 14.541 + .742 \times 50 = \underset{2.14}{51.64}$ for an estimated gain of

only .46 points less than the gain expected for a student with a low pretest score. Gains for Structure and Written Expression are thus quite uniform across the score range.

Again, the prediction lines have been added to the scatterplot.

The relationship of pretest, reliability test, and posttest for Reading Comprehension and Vocabulary is obtained with the request:

Scattergram B3 (20,70), C3 (20,70) WITH A3 (20,70)

Statistics ALL (Table B-20)

The resulting output is reproduced in Tables B-21 to B-24.

The equation for the baseline is

$\hat{B3} = 4.607 + .940\ A3$ and for the posttest

$\hat{C3} = 9.259 + .889\ A3.$

According to these formulas, a student with a Reading Comprehension and Vocabulary pretest score of 30 would have a predicted reliability test score of

$4.607 + .940 \times 30 = 32.29$ and posttest

$9.259 + .889 \times 30 = 35.93$ for an estimated gain of 3.14

A student with a pretest score of 50 would have a predicted reliability test score of

$4.607 + .940 \times 50 = 51.59$ and posttest

$9.259 + .889 \times 50 = 53.72$ for a gain of 2.13

Again, these estimated gains are not strikingly different across the score range.

After the balance of the analysis request cards, the deck ends with a FINISH card.

The computer system used for the sample analysis requires a card after FINISH, but this end-of-job signal may be different at another computer facility.

The complete listing of the deck for the sample run is given in Appendix A. It may be useful to punch these cards and to perform a test run to check that the procedures are compatible with your version of SPSS. If the sample run works but your real data analysis does not, check carefully for keypunch errors, missing punctuation, and cards out of order. Computers are ridiculously literal contraptions, and will not fill in an omitted comma in a set of instructions.

Interpreting Patterns of Change

If the reliability test is given within a week after the pretest, it is probably reasonable to assume that there has not been enough time for a significant change to result from instruction. This does not mean that individuals' scores are expected to be identical from pretest to reliability test: because of measurement error, neither test is a perfectly reliable indicator of true score, and the correlation between the tests, r_{12}, will be less than one because of this. In addition, small changes due to practice effects and increased comfort with the testing situation will take place even in brief intervals between test administrations. Thus, the group mean may well go up, and the test variance may change in the process of establishing our no-change baseline. The baseline is more accurately thought of as little influenced by the important sources of change--instructional programs--that we are studying.

In the case of our real data, the slope of the line predicting BTOT from ATOT, .969, is almost 1.0, and is greater than the correlation, .922. This is an indication that the variance of test B is greater than that of test A. In fact, the ratio of the standard deviation of B to that of A must be $\frac{.969}{.922} = 1.05$, (as can also be obtained from Table B-2) so that the variance has increased by $(1.05)^2$, a 10.25 percent increase from pretest to reliability test. Thus, although regression to the mean does take place by about 8 percent (1 - .922), it is almost offset by the 5 percent increase in standard deviation. This increase in spread of scores suggests that initially higher-scoring students benefitted more from the experience of taking the pretest, or learned more during the intervening week before the reliability test. For the total scores of this particular sample, using the no-change baseline will not yield conclusions very different from those obtained from using raw gain. The standard deviation of CTOT, the posttest, is only 97 percent that of the pretest, ATOT, and the slope of CTOT on ATOT is .841. Overall change shows a slight negative correlation with pretest score, and total score growth from a no-change baseline with slope very nearly 1.0 remains greater for those with initially low scores. The ratio of posttest slope to reliability test slope is $\frac{.841}{.969} = .868$.

Let us examine the subscores to determine if this pattern is the case for the separate parts of the test.

The Listening Comprehension subscores are denoted A1, B1, and C1. The correlation of B1 with A1 is .794, and the slope is .872. The variance of the Listening Comprehension section changes very little from pretest to B1, the reliability test, with the standard deviation decreasing by only about 1.5 percent. The slope of C1 vs. A1 is .650, and the correlation is .695. The standard deviation of C1 thus shows considerable further decrease, to only 93.5 percent that of A1, and this decrease indicates that change is correlated negatively with pretest score to a considerable extent. The ratio of posttest slope to reliability test

slope is relatively low: $\frac{.650}{.782} = .831$. This comparatively flat posttest slope yields high growth estimates for low pretest scorers, and low growth estimates for those with high pretest scores.

It appears that much of the apparent greater growth of low scorers on total test scores is due to this pattern for the Listening Comprehension subscores.

The Structure and Written Expression scores, A2, B2, and C2, exhibit a different pattern. The slope relating B2 to A2, .765, is almost identical to the slope of C2 vs. A2, .742. The correlation of B2 and A2, .760, is essentially the same as the slope, showing that no change in variance occurred between the pretest and the reliability test.

The ratio of the B2 and C2 slopes, .97, tells us that growth measured from the baseline, B2 to posttest C2, is essentially uniform across the score range for Structure and Written Expression, with initially high-scoring students gaining nearly exactly as much as do lower-scoring students when measurement error and practice effects have been taken into account. This occurs despite the decrease in standard deviation from pretest to posttest, the standard deviation of C2 being only 92.5 percent of the standard deviation of A2.

The only way that a decreasing standard deviation can be associated with constant gains from the baseline to the posttest is for these gains to be positively associated with pretest scores, but negatively associated with a component of the variance of the reliability test that is not related to either pretest or posttest. Indeed, the A2 - C2 correlation, .803, is higher than the A2 - B2 correlation, .760, suggesting that the "lost" variance in C2 was not related to pretest variance. One mechanism for such a result would be a short-term practice effect for some students on the reliability test, which washed out because of additional "test-wiseness" among all students by the time of the posttest.

The Reading and Vocabulary scores, A3, B3, and C3, show an almost 10 percent increase in standard deviation from pretest to reliability test, and a slope of .940, again approaching 1.0. Unlike the other subtests, posttest standard deviation remains about 6 percent greater than that of the pretest, rather than dropping to only about 93 percent of that value.

The correlation of A3 with C3, .834, is only slightly less than that of A3 with B3, .858, and the slope of C3 vs. A3, .889, is .95 times the slope of B3 vs. A3. Again, as with Structure and Written Expression, growth is almost uniform across the score scale, with high-scoring students gaining only slightly less than low scorers, after allowing for the effects of measurement error.

The general tendency for the reliability test to have greater variance than the posttest suggests that differential familiarization effects do take place in short-term retesting, temporarily adding variance that "washes out" over the longer term. Although such additional variance

can depress correlations between pretest and reliability test, it interferes much less with regression lines, affecting their standard error rather than their slope.

In any of these cases, the regression line predicting Test 2 score from Test 1 score determines a baseline expectation consisting of those average observed-score changes that are attributable to measurement error and to test practice effects. We now get on with teaching English, and administer a posttest at the point at which we wish to evaluate growth. We count as change due to instruction, neither raw gain from pretest, nor raw gain from the reliability test, but the difference between observed posttest score and the expected Test 2 score predicted for each pretest score by our regression equation, which gives a predicted score based on experience with these students for each pretest score. If no further change due to instruction has taken place since the reliability testing, we do not expect individual students to achieve scores identical to their scores on the reliability test (measurement error, again), but we do expect students in a given pretest score range to have posttest scores clustering around the same prediction line obtained from the reliability administration.

If we wished to be extremely conservative, we could choose to assume that any true changes observed from pretest to reliability test would have happened again without instruction, and double those changes to obtain an expectation for a third testing. This would be unreasonable, however, since test familiarization effects are not likely to operate strongly among those already familiar with the test. The major advantage of the reliability testing is to enable us to estimate reliability, and likely regression effects, in our particular student group. Unless our students are representative of all students who take TOEFL, reliability and probable regression effects are likely to differ from the TOEFL Manual statistics, based on samples of all candidates. However, if we consistently obtain similar baseline prediction equations over several semesters, and if the entering population in our program does not change in native language background or distribution of proficiency level, we can eventually dispense with a new reliability testing for each group and use the equation developed for previous groups, giving a new reliability testing only occasionally, to check the continuing validity of our local prediction equation.

```
SPSS BATCH SYSTEM                                                    08/10/81        PAGE    1

SPSS FOR OS/360; VERSION M; RELEASE 9.0; JUNE 10, 1981

                              CURRENT DOCUMENTATION FOR THE SPSS BATCH SYSTEM
ORDER FROM MCGRAW-HILL:  SPSS, 2ND ED. (PRINCIPAL TEXT)      ORDER FROM SPSS INC.:  SPSS STATISTICAL ALGORITHMS
                         SPSS UPDATE 7-9 (USE W/SPSS,2ND FOR REL. 7, 8, 9)        KEYWORDS: THE SPSS INC. NEWSLETTER
                         SPSS POCKET GUIDE, RELEASE 9
                         SPSS PRIMER (BRIEF INTRO TO SPSS)

DEFAULT SPACE ALLOCATION..     ALLOWS FOR..   102 TRANSFORMATIONS
WORKSPACE    71680 BYTES                     409 RECODE VALUES + LAG VARIABLES
TRANSPACE    10240 BYTES                    1641 IF/COMPUTE OPERATIONS

               1 RUN NAME     LANGUAGE GAIN ANALYSIS
               2 VARIABLE LIST  A1,A2,A3,ATOT,B1,B2,B3,BTOT,C1,C2,C3,CTOT
               3 INPUT MEDIUM   CARD
               4 INPUT FORMAT   FIXED(4X,3F2.0,1F3.0/4X,3F2.0,1F3.0/4X,3F2.0,1F3.0)

               ACCORDING TO YOUR INPUT FORMAT, VARIABLES ARE TO BE READ AS FOLLOWS

               VARIABLE  FORMAT  RECORD      COLUMNS

               A1        F 2.0       1      5-   6
               A2        F 2.0       1      7-   8
               A3        F 2.0       1      9-  10
               ATOT      F 3.0       1     11-  13
               B1        F 2.0       2      5-   6
               B2        F 2.0       2      7-   8
               B3        F 2.0       2      9-  10
               BTOT      F 3.0       2     11-  13
               C1        F 2.0       3      5-   6
               C2        F 2.0       3      7-   8
               C3        F 2.0       3      9-  10
               CTOT      F 3.0       3     11-  13

THE INPUT FORMAT PROVIDES FOR   12 VARIABLES.   12 WILL BE READ
IT PROVIDES FOR    3 RECORDS ('CARDS') PER CASE.  A MAXIMUM OF    13 'COLUMNS' ARE USED ON A RECORD.

               5 N OF CASES     98
               6 PEARSON CORR   A1 TO CTOT
               7 OPTIONS        5
               8 STATISTICS     1

 ***** PEARSON CORR PROBLEM REQUIRES     3168 BYTES WORKSPACE *****

               9 READ INPUT DATA
```

LANGUAGE GAIN ANALYSIS

FILE NONAME (CREATION DATE = 08/10/81)

VARIABLE	CASES	MEAN	STD DEV
A1	98	44.8469	6.9629
A2	98	37.6531	7.3528
A3	98	37.4082	6.6001
ATOT	98	399.7653	61.5730
B1	98	46.0510	6.8585
B2	98	40.0510	7.3979
B3	98	39.7551	7.2369
BTOT	98	419.5714	64.6636
C1	98	51.3061	6.5147
C2	98	42.4898	6.8009
C3	98	42.5204	7.0422
CTOT	98	455.3367	59.9381

B.111

FILE NONAME (CREATION DATE = 08/10/81)

PEARSON CORRELATION COEFFICIENTS

	A1	A2	A3	ATOT	B1	B2	B3	BTOT	C1	C2
A1	1.0000	0.6184**	0.6597**	0.8584**	0.7944**	0.6550**	0.6893**	0.7876**	0.6944**	0.5874**
A2	0.6184**	1.0000	0.7356**	0.8936**	0.6872**	0.7603**	0.7495**	0.8127**	0.6253**	0.8025**
A3	0.6597**	0.7356**	1.0000	0.8986**	0.7393**	0.6951**	0.8580**	0.8468**	0.6367**	0.7050**
ATOT	0.8584**	0.8936**	0.8986**	1.0000	0.8364**	0.7975**	0.8643**	0.9224**	0.7380**	0.7920**
B1	0.7944**	0.6872**	0.7393**	0.8364**	1.0000	0.6910**	0.7798**	0.9080**	0.8095**	0.6943**
B2	0.6550**	0.7603**	0.6951**	0.7975**	0.6910**	1.0000	0.6961**	0.8855**	0.6301**	0.7554**
B3	0.6893**	0.7495**	0.8580**	0.8643**	0.7798**	0.6961**	1.0000	0.9144**	0.7186**	0.7406**
BTOT	0.7876**	0.8127**	0.8468**	0.9224**	0.9080**	0.8855**	0.9144**	1.0000	0.7945**	0.8099**
C1	0.6944**	0.6253**	0.6367**	0.7380**	0.8095**	0.6301**	0.7186**	0.7945**	1.0000	0.6544**
C2	0.5874**	0.8025**	0.7050**	0.7920**	0.6943**	0.7554**	0.7406**	0.8099**	0.6544**	1.0000
C3	0.5882**	0.7135**	0.8343**	0.8034**	0.7173**	0.6424**	0.8930**	0.8319**	0.6830**	0.7202**
CTOT	0.6948**	0.7913**	0.8049**	0.8640**	0.8188**	0.7494**	0.8598**	0.8961**	0.8604**	0.8810**

* - SIGNIF. LE .01 ** - SIGNIF. LE .001 (99.0000 IS PRINTED IF A COEFFICIENT CANNOT BE COMPUTED)

B.12

FILE NONAME (CREATION DATE = 08/10/81)

PEARSON CORRELATION COEFFICIENTS

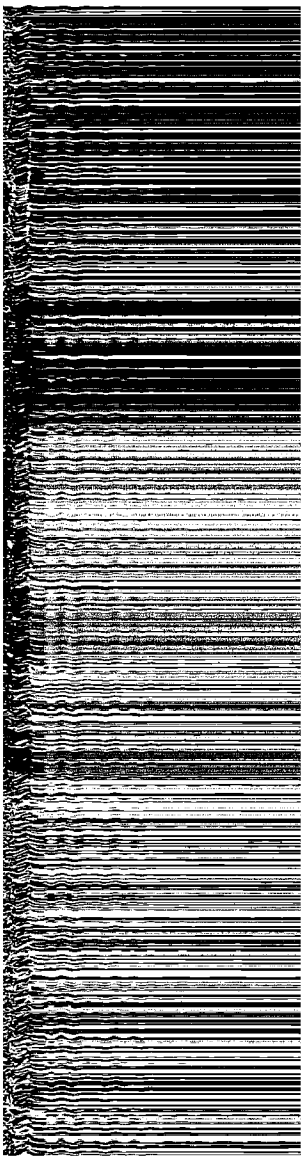
	C3	CTOT
A1	0.5882**	0.6948**
A2	0.7135**	0.7913**
A3	0.8343**	0.8049**
ATOT	0.8034**	0.8640**
B1	0.7173**	0.8188**
B2	0.6424**	0.7494**
B3	0.8930**	0.8598**
BTOT	0.8319**	0.8961**
C1	0.6830**	0.8604**
C2	0.7202**	0.8810**
C3	1.0000	0.8884**
CTOT	0.8884**	1.0000

* - SIGNIF. LE .01 ** - SIGNIF. LE .001 (99.0000 IS PRINTED IF A COEFFICIENT CANNOT BE COMPUTED)

CPU TIME REQUIRED.. 0.89 SECONDS

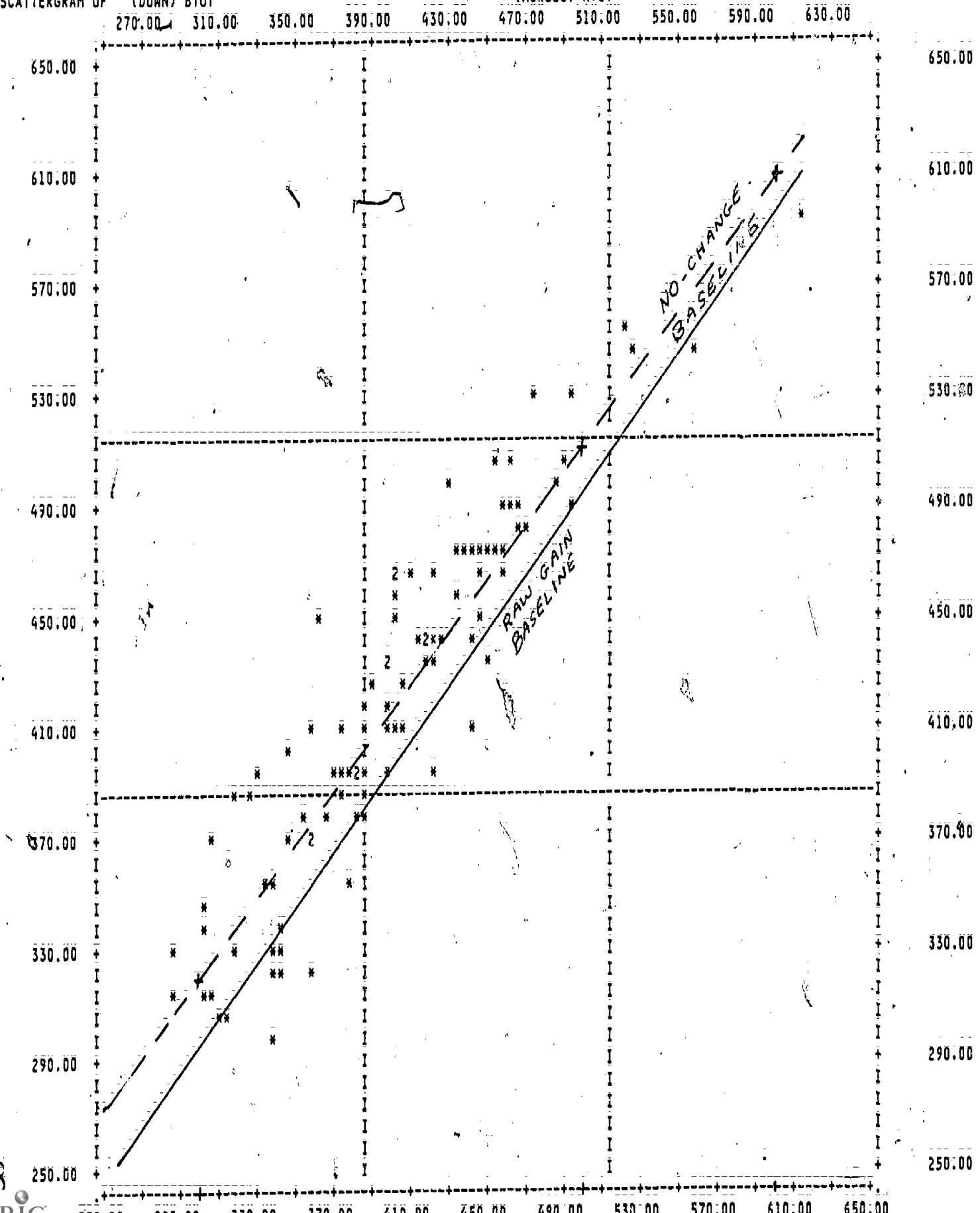
```
10 SCATTERGRAM   BTOT(250,650),CTOT(250,650) WITH ATOT(250,650)
11 STATISTICS    ALL
```

***** GIVEN WORKSPACE ALLOWS FOR 4478 CASES FOR SCATTERGRAM PROBLEM *****



LANGUAGE GAIN ANALYSIS 08/10/81 PAGE 6

FILE NONAME (CREATION DATE = 08/10/81)
SCATTERGRAM OF (DOWN) BTOT (ACROSS) ATOT



LANGUAGE GAIN ANALYSIS 08/10/81 PAGE 7

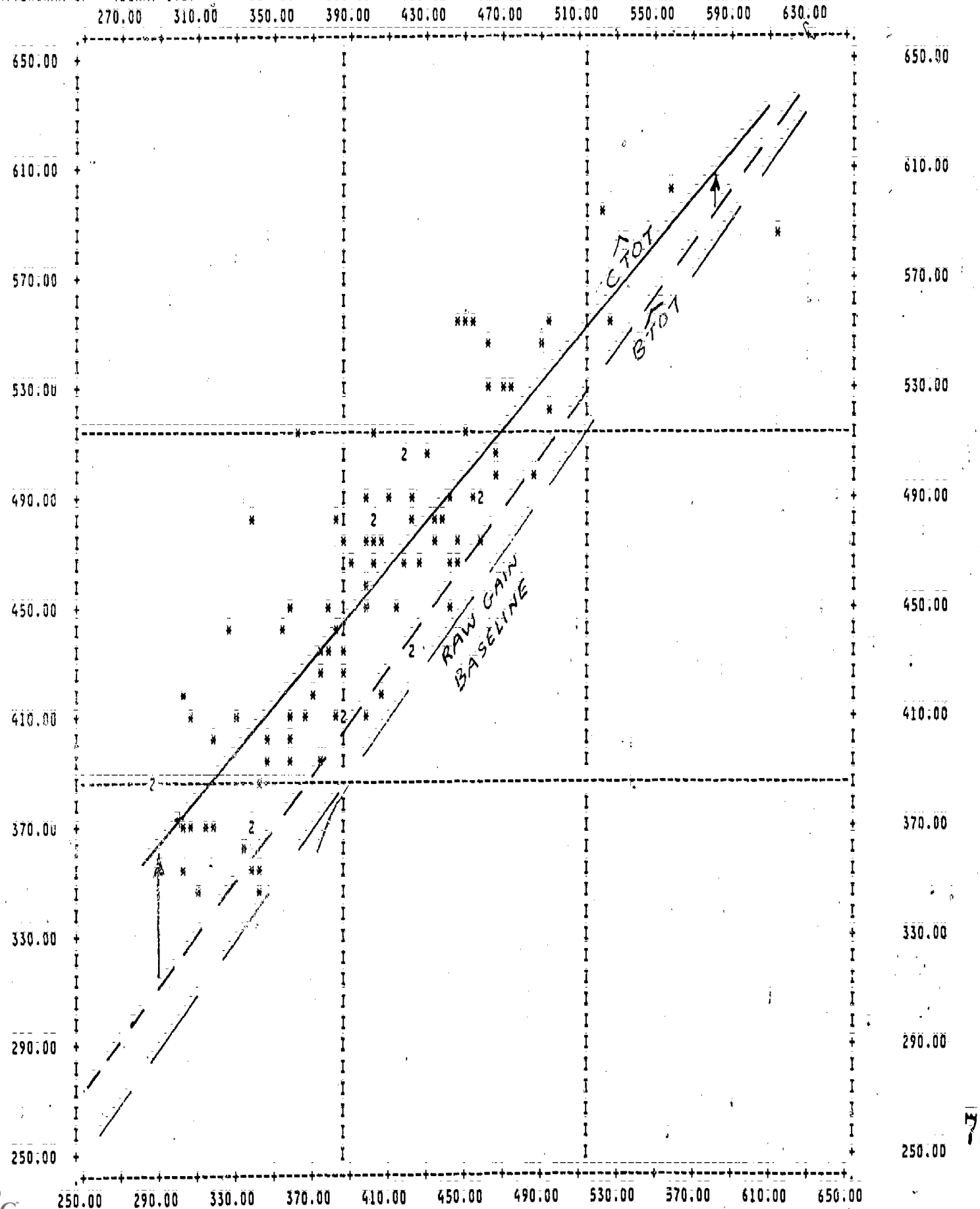
STATISTICS..

CORRELATION (R)-	0.92235	R SQUARED -	0.85073	SIGNIFICANCE -	0.00000
STD ERR OF EST -	25.11263	INTERCEPT (A) -	52.33932	SLOPE (B) -	0.96865
PLOTTED VALUES -	98	EXCLUDED VALUES-	0	MISSING VALUES -	0

'*********' IS PRINTED IF A COEFFICIENT CANNOT BE COMPUTED.

LANGUAGE GAIN ANALYSIS 08/10/81 PAGE 8

FILE NONAME (CREATION DATE = 08/10/81)
SCATTERGRAM OF (DOWN) CTOT (ACROSS) ATOT



STATISTICS..

CORRELATION (R)-	0.86402	R SQUARED -	0.74653	SIGNIFICANCE -	0.00000
STD ERR OF EST -	30.33321	INTERCEPT (A) -	119.10332	SLOPE (B) -	0.84108
PLOTTED VALUES -	98	EXCLUDED VALUES-	0	MISSING VALUES -	0

'********' IS PRINTED IF A COEFFICIENT CANNOT BE COMPUTED.

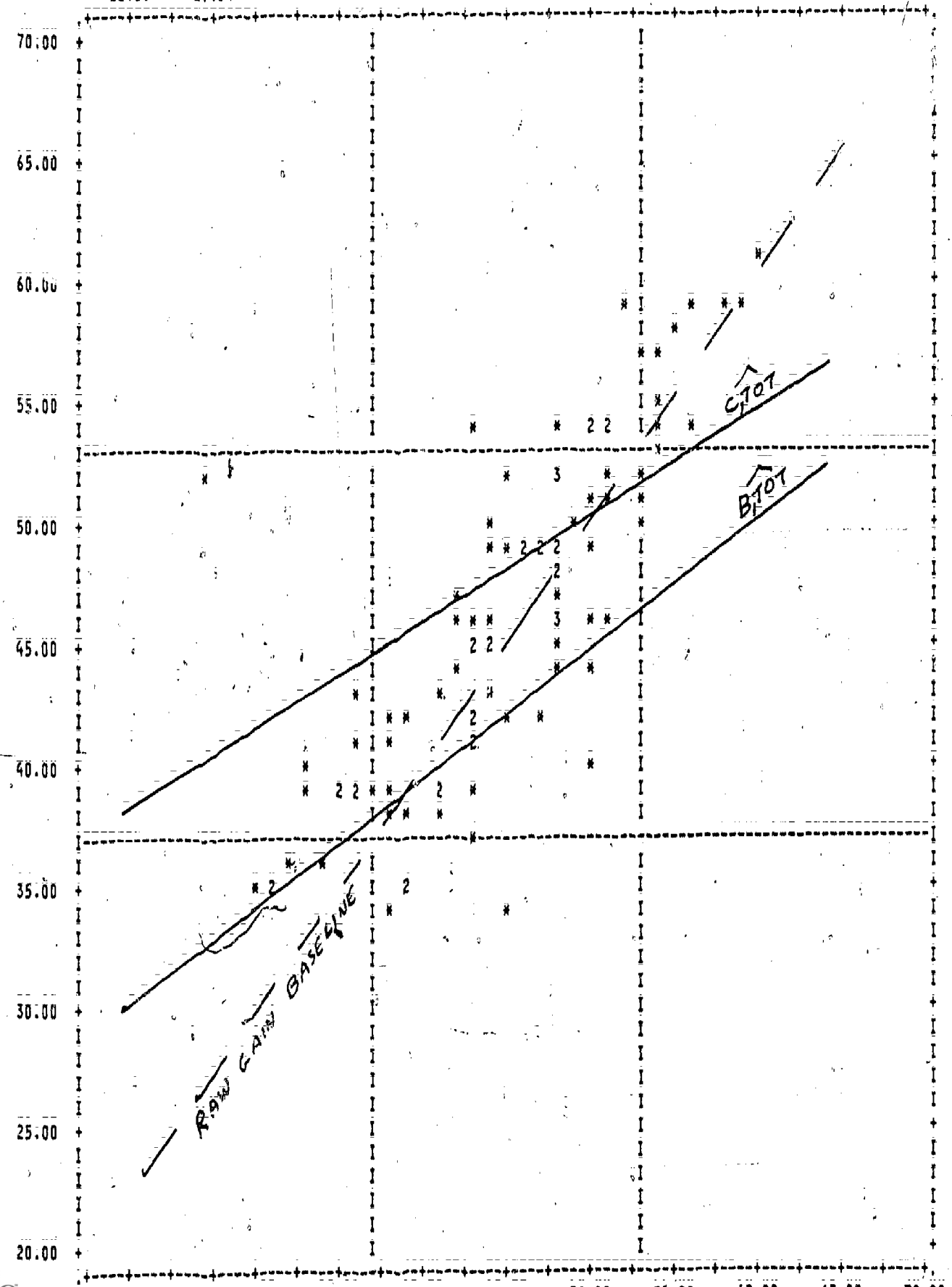
CPU TIME REQUIRED.. 0.73 SECONDS

```
12 SCATTERGRAM   B1(20,70),C1(20,70) WITH A1(20,70)
13 STATISTICS    ALL
```

***** GIVEN WORKSPACE ALLOWS FOR 4478 CASES FOR SCATTERGRAM PROBLEM *****

LANGUAGE GAIN ANALYSIS 08/10/81 PAGE 11

FILE NONAME (CREATION DATE = 08/10/81)
SCATTERGRAM OF (DOWN) B1 (ACROSS) A1



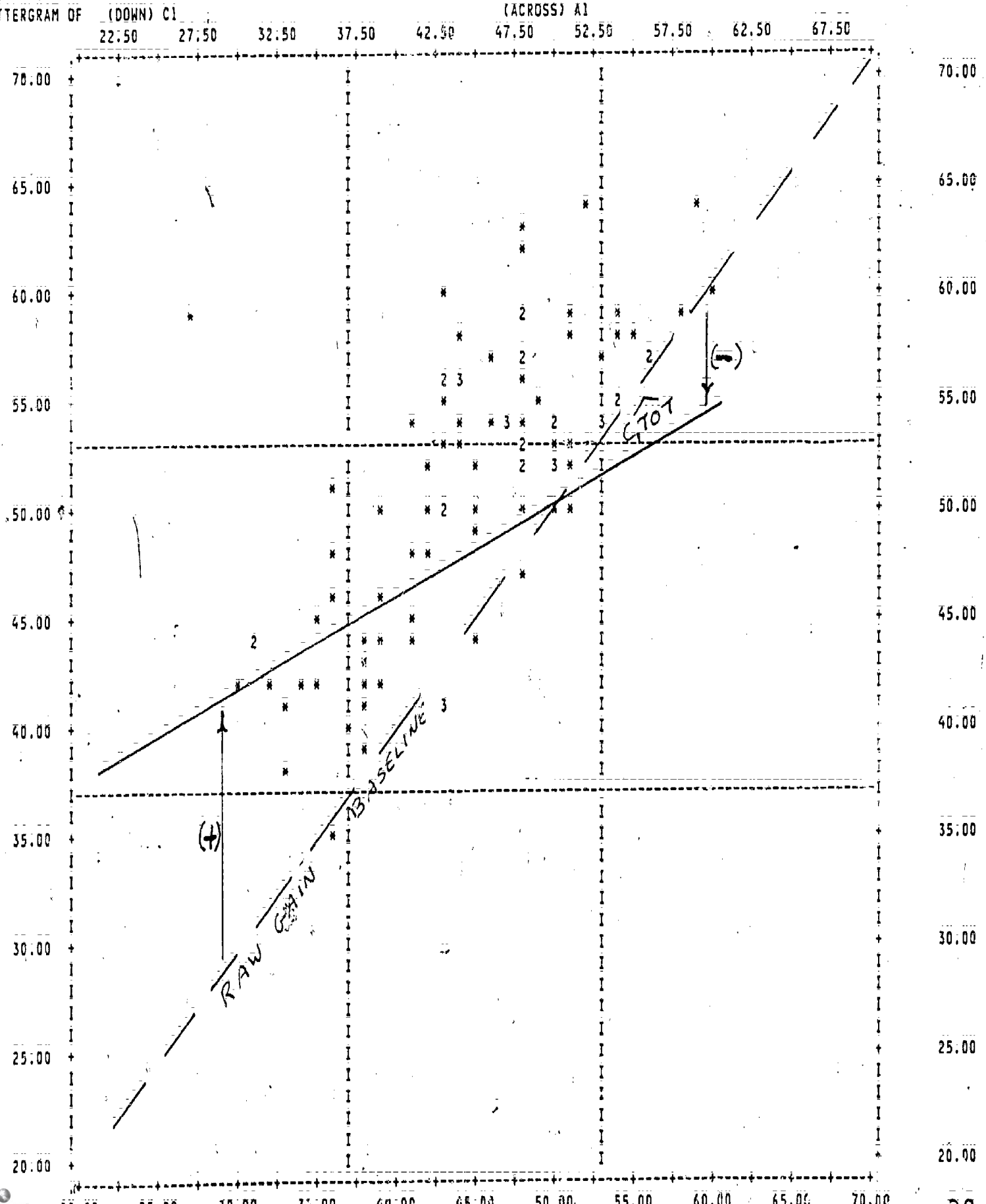
STATISTICS..

CORRELATION (R)-	0.79439	R SQUARED -	0.63105	SIGNIFICANCE -	0.00000
STD ERR OF EST -	4.18754	INTERCEPT (A) -	10.95925	SLOPE (B) -	0.78248
PLOTTED VALUES -	98	EXCLUDED VALUES-	0	MISSING VALUES -	0

'*********' IS PRINTED IF A COEFFICIENT CANNOT BE COMPUTED.

LANGUAGE GAIN ANALYSIS 08/10/81 PAGE 13

FILE NONAME (CREATION DATE = 08/10/81)
SCATTERGRAM OF (DOWN) C1 (ACROSS) A1



LANGUAGE GAIN ANALYSIS 08/10/81 PAGE 14

STATISTICS:

CORRELATION (R)-	0.69445	R SQUARED	-	0.48226	SIGNIFICANCE	-	0.00000
STD ERR OF EST -	4.71195	INTERCEPT (A)	-	22.13673	SLOPE (B)	-	0.64975
PLOTTED VALUES -	93	EXCLUDED VALUES-		0	MISSING VALUES	-	0

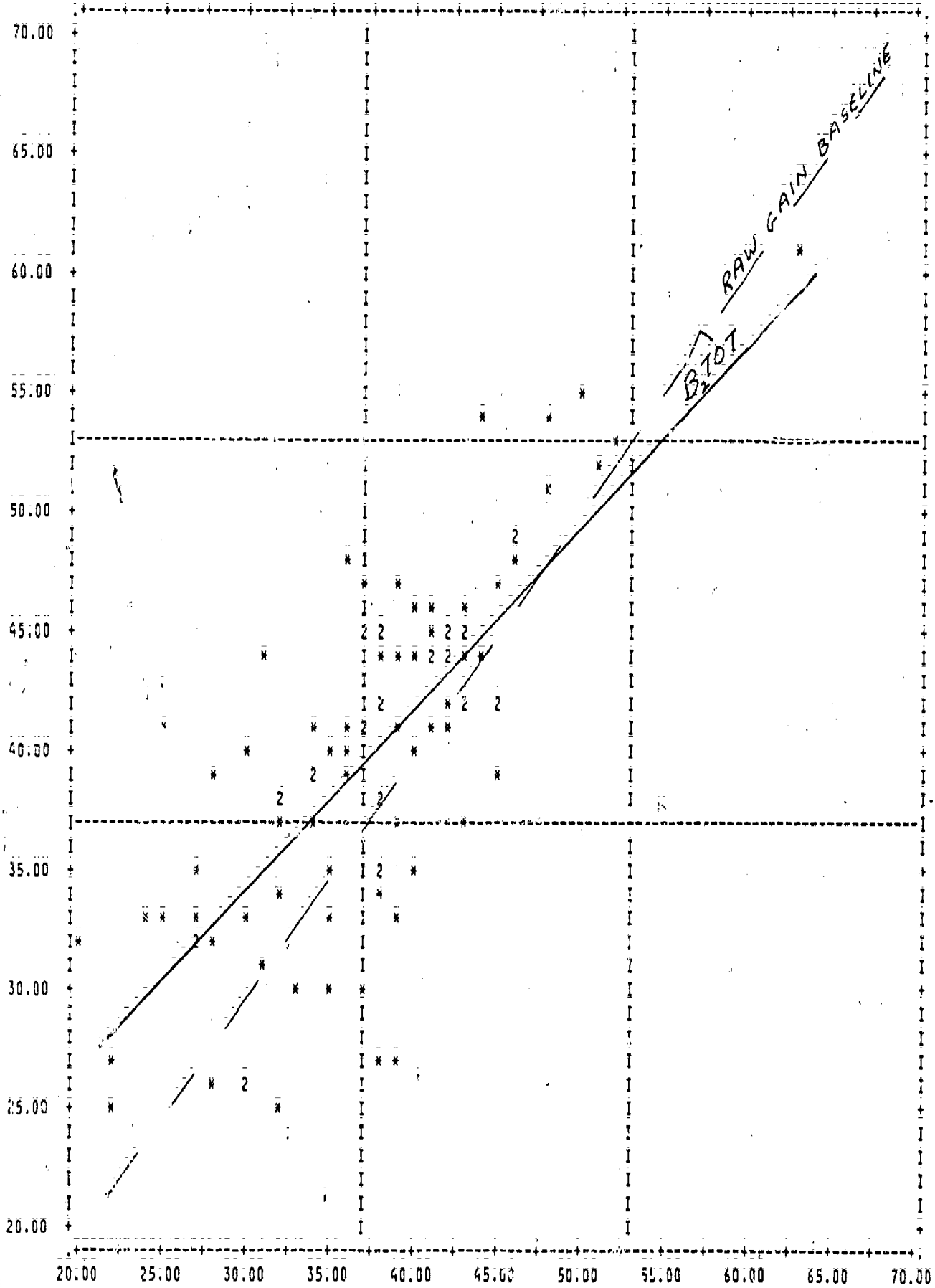
'*********' IS PRINTED IF A COEFFICIENT CANNOT BE COMPUTED.

CPU TIME REQUIRED.. 0.71 SECONDS

14 SCATTERGRAM B2(20,70),C2(20,70) WITH A2(20,70)
15 STATISTICS ALL

***** GIVEN WORKSPACE ALLOWS FOR 4478 CASES FOR SCATTERGRAM PROBLEM *****

FILE NONAME (CREATION DATE = 08/10/81)
SCATTERGRAM OF (DOWN) B2 (ACROSS) A2



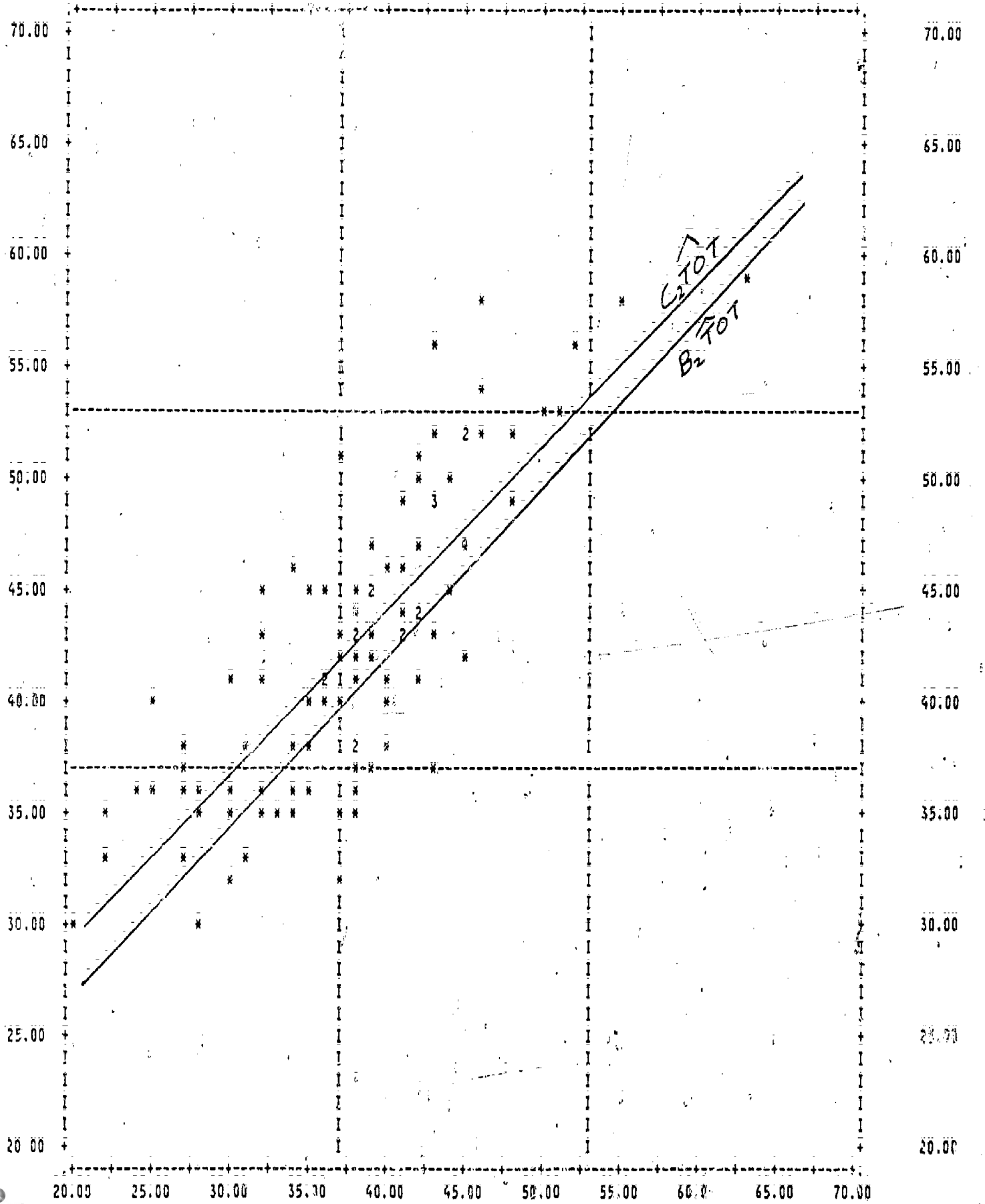
LANGUAGE GAIN ANALYSIS

STATISTICS:

CORRELATION (R)-	0.76032	R SQUARED -	0.57809	SIGNIFICANCE -	0.00000
STD ERR OF EST -	4.83028	INTERCEPT (A) -	11.24701	SLOPE (B) -	0.76498
PLOTTED VALUES -	98	EXCLUDED VALUES-	0	MISSING VALUES -	0

'*********' IS PRINTED IF A COEFFICIENT CANNOT BE COMPUTED.

FILE NONAME (CREATION DATE = 08/10/81)
SCATTERGRAM OF (DOWN) C2 (ACROSS) A2

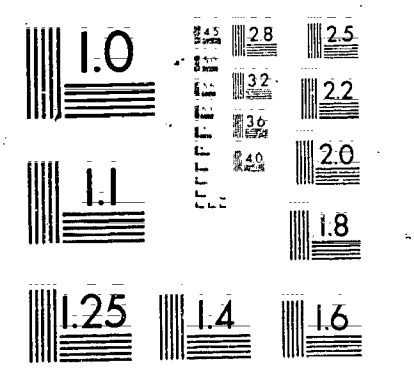


STATISTICS..

CORRELATION (R)-	0.80251	R SQUARED -	0.64403	SIGNIFICANCE -	0.00000
STD ERR OF EST -	4.07873	INTERCEPT (A) -	14.54079	SLOPE (B) -	0.74228
PLOTTED VALUES -	98	EXCLUDED VALUES-	0	MISSING VALUES -	0

'*********' IS PRINTED IF A COEFFICIENT CANNOT BE COMPUTED.

B-28



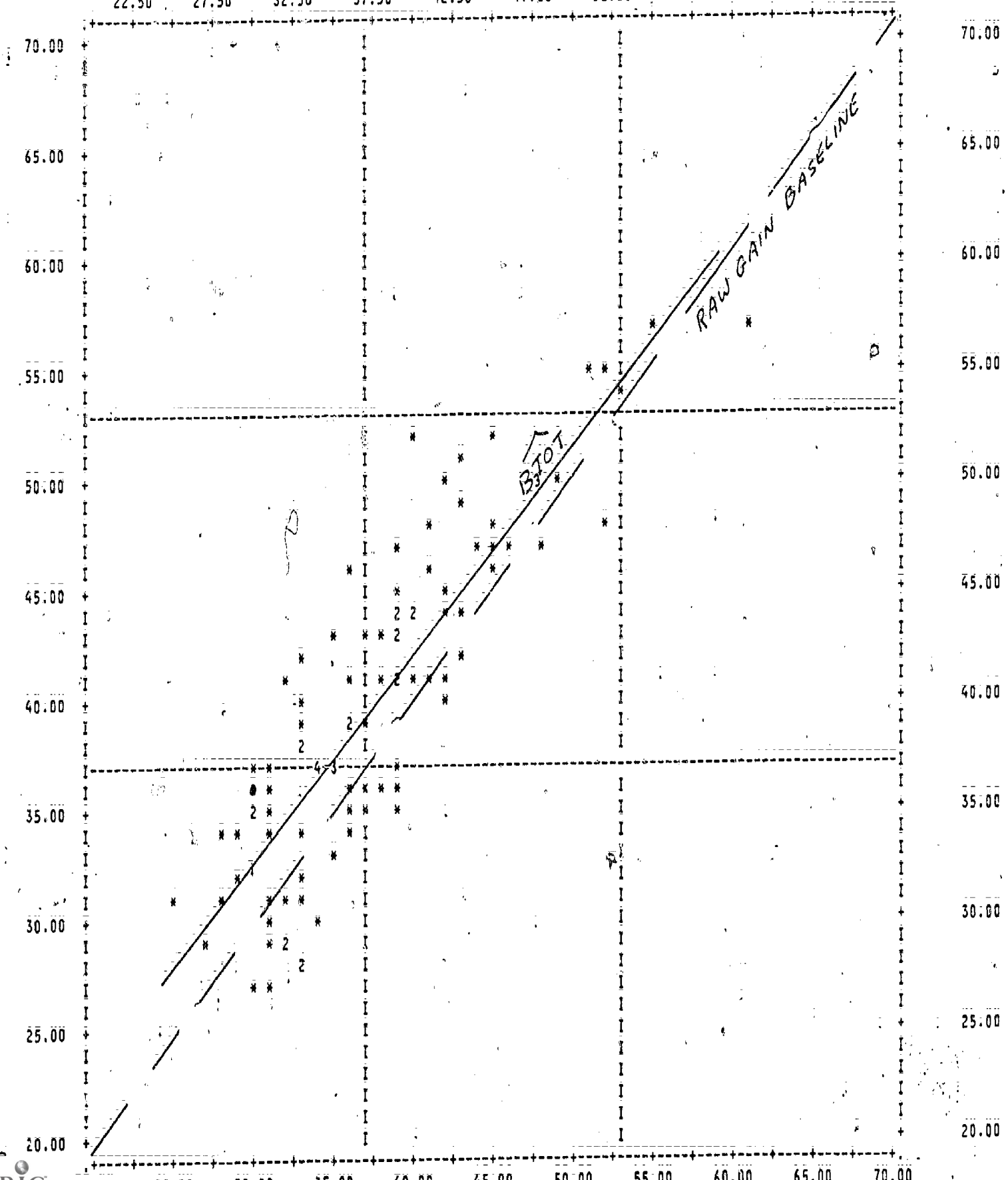
CPU TIME REQUIRED.. 0.73 SECONDS

16 SCATTERGRAM B3(20,70),C3(20,70) WITH A3(20,70)
17 STATISTICS ALL

***** GIVEN WORKSPACE ALLOWS FOR 4478 CASES FOR SCATTERGRAM PROBLEM *****

LANGUAGE GAIN ANALYSIS 08/10/81 PAGE 21

FILE NONAME (CREATION DATE = 08/10/81)
SCATTERGRAM OF (DOWN) B3 (ACROSS) A3



LANGUAGE GAIN ANALYSIS 08/10/81 PAGE 22

STATISTICS..

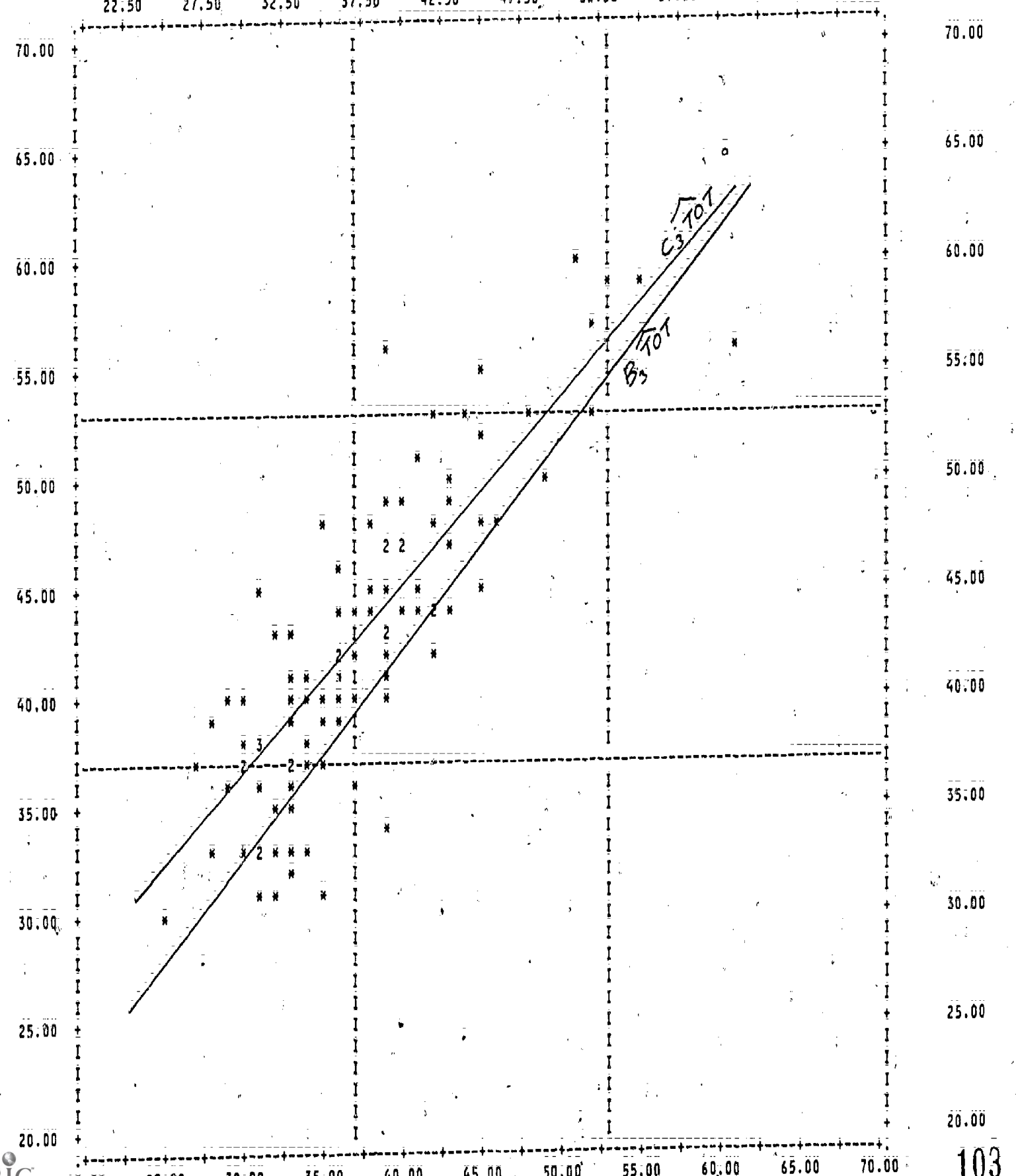
CORRELATION (R)-	0.85795	R SQUARED	-	0.73608	SIGNIFICANCE	-	0.00000
STD ERR OF EST -	3.73713	INTERCEPT (A)	-	4.60677	SLOPE (B)	-	0.93959
PLOTTED VALUES -	98	EXCLUDED VALUES-		0	MISSING VALUES	-	0

'********' IS PRINTED IF A COEFFICIENT CANNOT BE COMPUTED.

LANGUAGE GAIN ANALYSIS

FILE NONAME (CREATION DATE = 08/10/81)

SCATTERGRAM OF (DOWN) C3 (ACROSS) A3



LANGUAGE GAIN ANALYSIS 08/10/81

STATISTICS..

CORRELATION (R)-	0.83435	R SQUARED -	0.69614	SIGNIFICANCE -	0.00000
STD ERR OF EST -	3.90209	INTERCEPT (A) -	9.25863	SLOPE (B) -	0.88916
PLOTTED VALUES -	98	EXCLUDED VALUES-	0	MISSING VALUES -	0

'*********' IS PRINTED IF A COEFFICIENT CANNOT BE COMPUTED.

CPU TIME REQUIRED.. 0.73 SECONDS

18 FINISH

NORMAL END OF JOB.
18 CONTROL CARDS WERE PROCESSED.
0 ERRORS WERE DETECTED.

B.34

Appendix C

Comparing Two Groups

On some occasions, we wish to compare two or more groups to determine if the growth rate for students of a particular language or educational background is typical of that for the total group, or to compare growth of different groups across semesters or across alternative ESL curricula.

This appendix presents two approaches to this problem. The first approach is simply to redo, for the subgroup, the regression analysis described in Appendix B and to compare the resulting equations and graphs with those of the total group. No tests of statistical significance are employed in this comparison, since the object is to determine whether any observed difference is large enough to be of practical significance to the program, rather than to assess probability. If tests of statistical significance are desired, the second approach may be employed. This method, analysis of covariance (ANCOVA), is also based on regression and yields an estimate of the difference between two groups at the mean pretest score, and of the probability that a difference this large could occur by chance.

Comparing Regression Lines

To illustrate this method, we use the subgroup of 30 returning students from the group of 98 ESL students discussed previously. Table C-1 shows the first page of the regression output, giving the control cards, format, and first analysis request. The 30 cases representing the returning students were separated from the total data deck for this run, so the only change from the control cards for the total group is in the number of cases, N = 30 cards. With a larger data set, it would be more convenient to change the input format to read the variable RET in column 79, and to use the SPSS "SELECT IF RET = 1" option to read the entire deck but to process only these 30 cases.

Table C-2 gives the means and standard deviations of the subtests and total tests. We see that the mean of pretest ATOT, 409.93, is about ten points higher than that of the total group and the mean of posttest CTOT, 458.5, is about three points higher than that of the total group. For this analysis, we ignore the reliability test B, and focus on comparability of changes from pretest to posttest.

Table C-3 shows the graph of posttest on pretest, and Table C-4 gives the intercept, 113.63, and slope, .841, of the regression equation. Comparing this equation,

$$\widehat{CTOT} = 113.63 + .841 \times ATOT$$

with that for the total group,

$$\widehat{CTOT} = 119.10 + .841 \times ATOT.$$

We see that the slopes are identical, but that for a given pretest total score, the returning students gain, on the average, about 5.5 points less than do all students. This does not appear to be a large enough difference to be of practical significance. If we were to do a separate analysis on the 68 new students, we would of course find that their intercept was a bit higher than 119, since the total group is the combination of the returning students and the new students. However, even a difference of eight total points between the two groups is not large.

Moving to the Listening Comprension test, Table C-5 shows the graph and Table C-6 the constants for the regression of C1 on A1 for returning students. The equation,

$\hat{C1} = 13.40 + .799 \times A1$, appears different from that for the total group:

$\hat{C1} = 22.1T + .650 \times A1.$

For a pretest Listening Comprehension score of 30, the returning group would have a predicted posttest score of 37.4, and for a pretest score of 50, a posttest score of 53.4. For the same two pretest scores, the total group's predicted posttest scores are 41.7 and 54.7, respectively.

Although a five-point difference on total scores is only one-eightieth of the mean score, and is not educationally significant, a four-point difference on the Listening Comprehension subtest for low-scoring students is about one-twelfth of the mean, and has considerable significance. It suggests that students who remain low on Listening Comprehension after a semester of instruction continue to progress at a lower than average rate in the following semester. This "sorting" effect on rate of learning of previous instruction helps to explain the higher slope for the returning group: those who learned more in the first semester, and who thus had higher pretest scores for the second semester, continue to progress at a faster than average rate, but still do not show as large a raw gain as new students, unless their pretest scores are 60 or higher.

Tables C-7 and C-8 give the graph and statistics for the prediction of Structure and Written Expression posttest scores from returning students' pretest scores. The equation,

$\hat{C2} = 9.93 + .875 \times A_2$, again has a higher slope and lower intercept than does the equation for the total group:

$\hat{C_2} = 11.25 + .765 \times A2.$

The difference is qualitatively unlike that for Listening Comprehension, however. Returning students with pretest scores of 30 and 50 would be predicted to achieve posttest scores of 36.2 and 53.7, respectively.

New students with the same pretest scores would be expected to reach average posttest scores of 34.2 and 49.5.

In this case, returning students gain from two to four more raw-score points than do new students, partly offsetting the lesser gain observed for Listening Comprehension scores.

Finally, the Reading Comprehension and Vocabulary scores and constants for the returning students are given in Tables C-9 and C-10. The resulting prediction equation is

$\hat{C3} = 8.17 + .903 \times A3$. This is almost identical to the equation for the total group:

$\hat{C3} = 9.26 + .889 \times A3$.

Returning students with Reading Comprehension and Vocabulary pretest scores of 30 and 50 would have predicted scores of 35.3 and 53.3, respectively, as compared to predicted scores of 35.9 and 53.7 for new students with the same pretest scores.

Comparison of the pretest-posttest relationships for returning students with those for new students has revealed that although slopes are identical for total test scores, and the lower total raw gain for returning students is not practically important, this total test-score pattern holds only for the Reading Comprehension and Vocabulary subtest. In Listening Comprehension, returning students with a given pretest score gained substantially less than did new students with the same score, while in Structure and Written Expression, returning students gained more than did new students.

If statistical tests of significance are desired for these differences, or for differences arising from the comparison of data from two or more different curricula or from two or more different semesters, the SPSS ANCOVA analysis offers a convenient method. ANCOVA uses one or more predictors (covariates) in regression equations to explain as much of the posttest variance as possible. The "leftover" or residual variance that cannot be explained by such covariates as pretest score, years of language study, or language group is then subjected to traditional analysis of variance. This procedure compares the variance among group means with the residual variation within groups to estimate the probability that observed group differences could have occurred by chance. The resulting "F" statistic has a probability distribution that depends on the number of groups and on the number of individuals within groups. An important assumption of the classical analysis of variance is that regression lines are parallel within groups. Although we have seen that the observed regression lines for new and returning students are not strictly parallel for subtests 1 and 2, the lack of parallelism is not significant and does not violate the parallelism assumption seriously.

Table C-11 shows the first page of output of an SPSS ANCOVA analysis using two groups: the 68 new students of our earlier analyses, and the 30 returning students.

The input format card has been changed to read column 79 of card 1, in which returning students are coded with the number 1. The variable list card has been changed by adding a variable "RET," which is 1 for returning students and 0 otherwise.

The first analysis request card,

ANOVA CTOT BY RET(0,1) WITH ATOT,

asks for an analysis of covariance with total posttest score CTOT as the dependent variable, RET as the group identification code, and pretest ATOT as the covariate. Table C-12 shows the CTOT means and numbers of the total sample and of each subgroup. Table C-13 gives the analysis of covariance results for the total test scores.

Although the covariate, ATOT, predicts a highly significant proportion of the variance of the total posttest score ($F = 283.88$, probability of this large a value by chance less than .0005), the main effect of returning status predicts very little of the posttest variance ($F = 1.39$, a value that could be observed by chance in almost one in four cases). The total variance explained is significant, but only because of the contribution of the covariate. Thus the statistical test confirms the judgment based on the comparison of regression lines: returning status does not significantly influence language growth predictions in this sample. Table C-14 gives the estimated unadjusted posttest differences, with the new group 1.4 points below the grand mean and the returning group 3.16 points above the grand mean, and the adjusted contrasts after taking the pretest into account, reversed to show a 5.46-point negative weighted effect for returning students and a 2.41-point positive weighted effect for new students. These effects are weighted by the number of cases to sum to zero ($68 \times 2.41 - 30 \times 5.46 = 0$).

As we have noted, this effect is not statistically significant.

Analysis of covariance is not restricted to a single covariate. Table C-15 gives the control cards for analysis of covariance using both the pretest and the reliability test as predictors. One could also use years of prior English study or a code for native language group (e.g., 0 for Indo-European, 1 for non-Indo-European) as additional predictors. One could not use months of English study in the U.S. as a covariate, since it would be confounded with new/returning status and would explain away the very effect that we wish to study.

Table C-16 repeats the means for the two groups, and Table C-17 gives the analysis of covariance table. We note that both covariates contribute significantly to predicting the posttest, with BTOT, being closer in time to the posttest, contributing more unique predictive power than does ATOT,

although their common component, F = 203.57, carries the burden of the prediction. The addition of BTOT as a covariate effectively wipes out any effect after the first week of returning status (F = 0, probability = 1). The contrasts in Table C-18 show that the adjusted mean differences have dropped to a negligible .13 - (-.30) = .43 points.

Table C-19 gives the control cards for the single-covariate analysis of the Listening Comprehension subtest, Table C-20 the means, and Table C-21 the analysis of variance table. Here again, the statistical test partly confirms our intuitive judgment. The effect of returning status on Listening Comprehension is statistically significant (p = .015). Table C-22 shows a covariance-adjusted effect of .79 - (-1.79) = 2.58 points in favor of new students. Adding the reliability test B1 as an additional covariate, however (Tables C-23 to C-26), makes the effect of new vs. returning status drop to statistical insignificance (p = .279), with an adjusted effect (Table C-26) of only .98.

The analyses of the Structure and Written Expression subtest comprise Tables C-27 to C-34. Using A2 as a covariate, the effect of group (returning vs. new) is not significant (F = .297, p = .587), and the estimated effect size is only .59 points (Table C-26). Adding B2, the reliability test, as an additional covariate drops the value of F to .052 (p = .819), and reduces the estimated effect to a negligible .19.

Tables C-35 to C-42 give the analyses for the Reading Comprehension and Vocabulary subtests. As was noted by comparing regression lines, differences are slight, and neither the analysis with pretest only as covariate (F = .888, p = .348) nor that with both pretest and reliability tests as covariates (F = .092, p = .763) approaches statistical significance.

The analysis of covariance thus offers a convenient test of group differences, with considerable increase in power afforded by using covariates to remove what would otherwise have been "error" variance. It should be kept in mind that the analysis of covariance presupposes random assignment to groups, however, and is not capable of correcting for preexisting differences among groups selected on some ability-correlated criterion. In cases in which the assumption of parallel within-group regression lines is violated, generalizations of analysis of covariance, which fit a group-by-pretest interaction to the data (Ragosa, 1981), may be considered.

In comparing different groups, it is important to check for differential attrition. If subjects have dropped out of both groups before posttest, it is necessary to check that the pretest scores of those who left each group are comparable. If they are not, some cause of dropping out, linked to test scores, may have been operating, and any difference in outcome may be attributable to this differential dropout rather than to some positive characteristic of the program. The classical example of this is the teacher who says, "you, you, and you, stay home tomorrow," on the day before the posttest, but more subtle influences may operate to produce a similar result.

Concluding Note

It should be stressed that we have not attempted to estimate differing true gain scores for individual students with the same pretest scores. The difficulties inherent in that task can (and have) filled a book (Harris, 1963).

Rather, we have followed the recommendations of Cronbach and Furby (1970), who point out that the correlational question, "What kinds of individuals grow more?" can be answered without estimating true gain scores for individuals, but is best approached by studying predicted scores.

Neither did we adopt the point of view that since measurement scales may be arbitrarily stretched, only changes uncorrelated with pretest ("structural changes") qualify as "real" change. For example, if every student were to gain 10 percent from his or her pretest level, posttest scores would correlate perfectly with pretest scores, but we would still claim that change had taken place, and that initially high-scoring students had gained more than had initially low-scoring students. This interpretation amounts to assuming that units of the TOEFL scale are meaningful to users in behavioral terms.

Good luck with your analysis, and please communicate any problems or suggestions for clarification to the author.

References

Cronbach, L. J., & Furby, L. How should we measure "change": Or should we? Psychological Bulletin, 1970, 74, 63-80.

Harris, C. W. Problems in measuring change. Madison: The University of Wisconsin Press, 1963.

Ragosa, D. Comparing nonparallel regression lines. Psychological Bulletin, 1980, 88, 307-321.

SPSS BATCH SYSTEM 08/11/81 PAGE 1

SPSS FOR OS/360, VERSION M, RELEASE 9.0, JUNE 10, 1981

CURRENT DOCUMENTATION FOR THE SPSS BATCH SYSTEM

ORDER FROM MCGRAW-HILL: SPSS, 2ND ED. (PRINCIPAL TEXT)
SPSS UPDATE 7-9 (USE W/SPSS,2ND FOR REL. 7, 8, 9)
SPSS POCKET GUIDE, RELEASE 9
SPSS PRIMER (BRIEF INTRO TO SPSS)

ORDER FROM SPSS INC.: SPSS STATISTICAL ALGORITHMS
KEYWORDS: THE SPSS INC. NEWSLETTER

DEFAULT SPACE ALLOCATION: ALLOWS FOR: 102 TRANSFORMATIONS
WORKSPACE 71680 BYTES 409 RECODE VALUES + LAG VARIABLES
TRANSPACE 10240 BYTES 1641 IF/COMPUTE OPERATIONS

```
   1 RUN NAME       LANGUAGE GAIN ANALYSIS
   2 VARIABLE LIST  A1,A2,A3,ATOT,B1,B2,B3,BTOT,C1,C2,C3,CTOT
   3 INPUT MEDIUM   CARD
   4 INPUT FORMAT   FIXED(4X,3F2.0,1F3.0/4X,3F2.0,1F3.0/4X,3F2.0,1F3.0)
```

ACCORDING TO YOUR INPUT FORMAT, VARIABLES ARE TO BE READ AS FOLLOWS

VARIABLE	FORMAT	RECORD	COLUMNS
A1	F 2.0	1	5- 6
A2	F 2.0	1	7- 8
A3	F 2.0	1	9- 10
ATOT	F 3.0	1	11- 13
B1	F 2.0	2	5- 6
B2	F 2.0	2	7- 8
B3	F 2.0	2	9- 10
BTOT	F 3.0	2	11- 13
C1	F 2.0	3	5- 6
C2	F 2.0	3	7- 8
C3	F 2.0	3	9- 10
CTOT	F 3.0	3	11- 13

THE INPUT FORMAT PROVIDES FOR 12 VARIABLES. 12 WILL BE READ
IT PROVIDES FOR 3 RECORDS ('CARDS') PER CASE. A MAXIMUM OF 13 'COLUMNS' ARE USED ON A RECORD.

```
   5 N OF CASES     30
   6 PEARSON CORR   A1 TO CTOT
   7 OPTIONS        5
   8 STATISTICS     1
```

***** PEARSON CORR PROBLEM REQUIRES 3168 BYTES WORKSPACE *****

```
   9 READ INPUT DATA
```

C-7

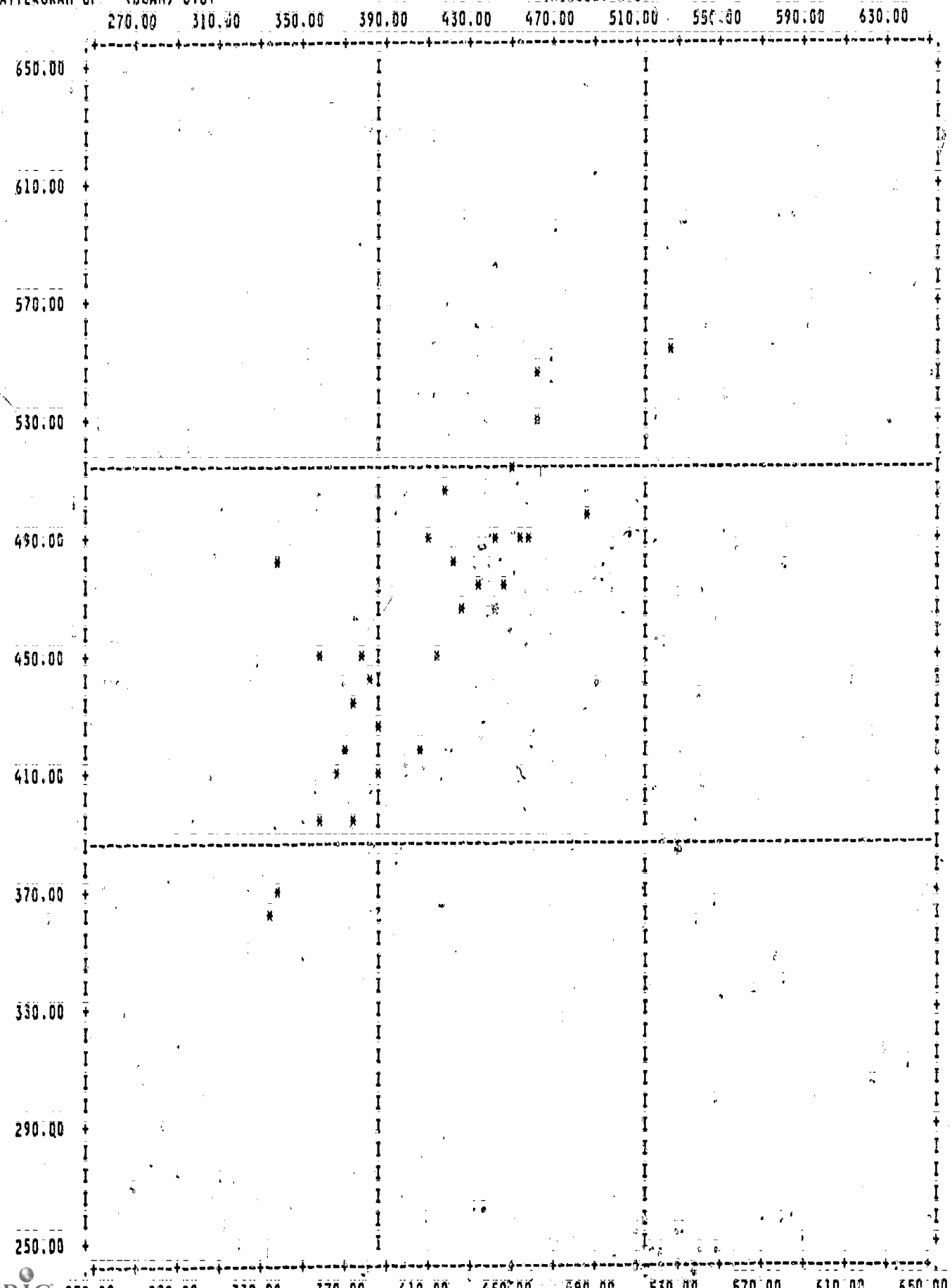
FILE NONAME (CREATION DATE = 08/11/81).

VARIABLE	CASES	MEAN	STD DEV
A1	30	47.4667	5.2110
A2	30	37.3000	5.0182
A3	30	38.2000	6.3594
ATOT	30	409.9333	48.0423
B1	30	46.4667	6.9765
B2	30	40.4333	5.6793
B3	30	39.4333	7.0987
BTOT	30	421.1333	57.9784
C1	30	51.3333	5.3584
C2	30	42.5667	5.9982
C3	30	42.6667	6.3698
CTOT	30	458.5000	48.9008

LANGUAGE GAIN ANALYSIS 08/11/81 PAGE 8

FILE NONAME (CREATION DATE = 08/11/81)

SCATTERGRAM OF (DOWN) CTOT (ACROSS) ATOT



LANGUAGE GAIN ANALYSIS 08/11/81 PAGE 9

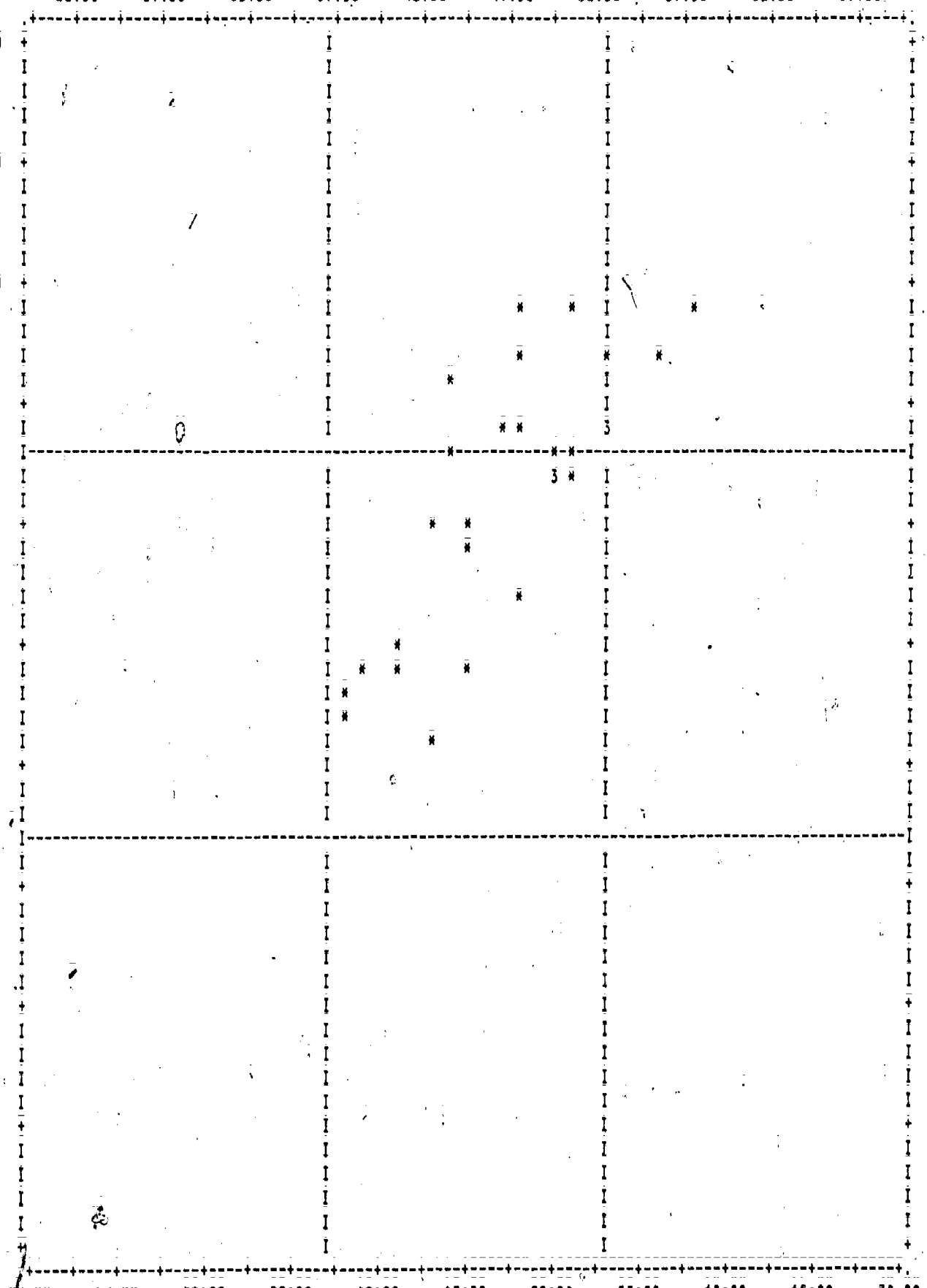
STATISTICS..

CORRELATION (R)-	0.82651	R SQUARED -	0.68312	SIGNIFICANCE -	0.00000
STD ERR OF EST -	28.01475	INTERCEPT (A) -	113.63204	SLOPE (B) -	0.84128
PLOTTED VALUES -	30	EXCLUDED VALUES-	0	MISSING VALUES -	0

'*********' IS PRINTED IF A COEFFICIENT CANNOT BE COMPUTED.

LANGUAGE GAIN ANALYSIS 08/11/81 PAGE 13

FILE NONAME (CREATION DATE = 08/11/81)
SCATTERGRAM OF (DOWN) C1 (ACROSS) A1



LANGUAGE GAIN ANALYSIS

STATISTICS..

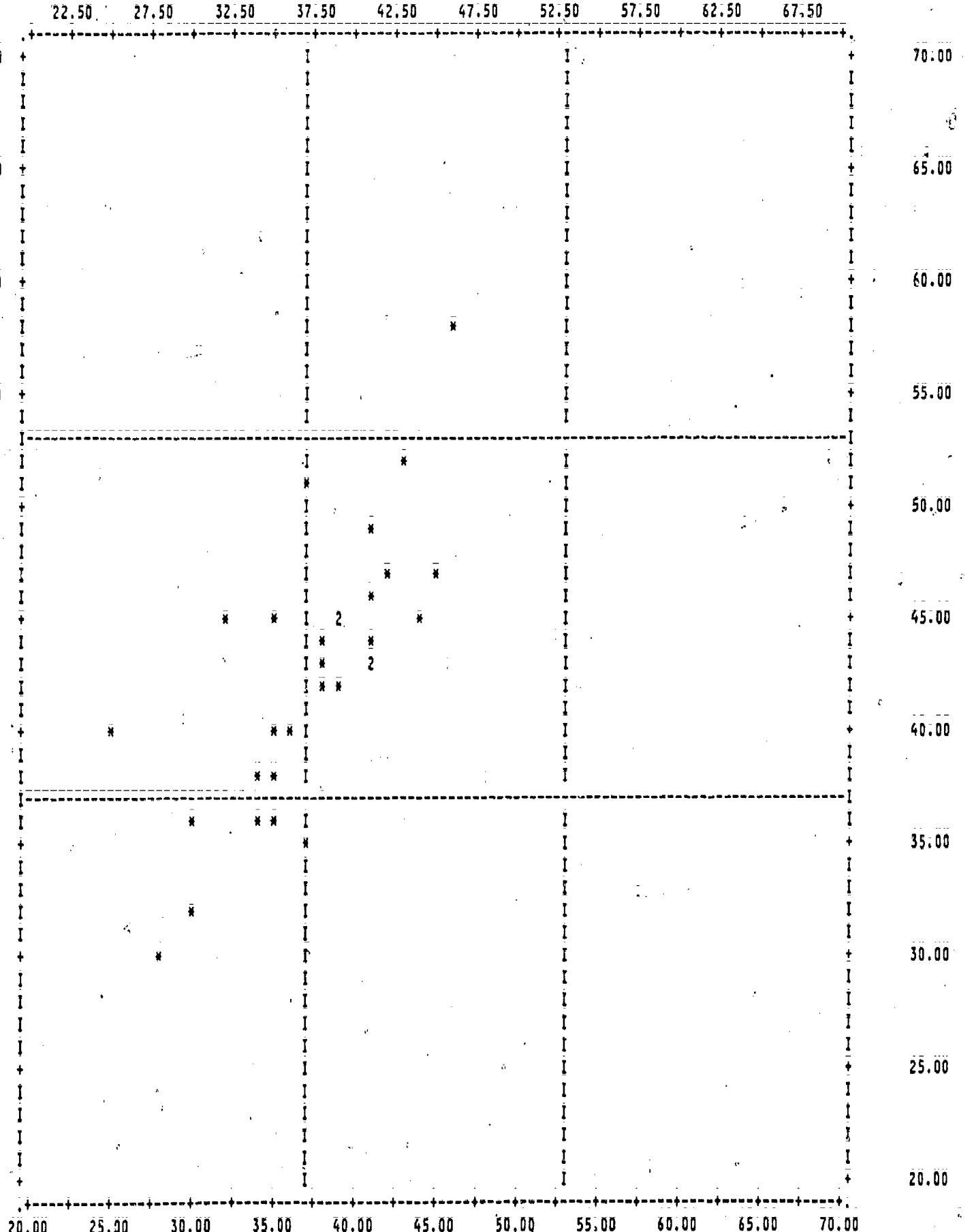
CORRELATION (R)-	0.77719	R SQUARED -	0.60403	SIGNIFICANCE -	0.00000
STD ERR OF EST -	3.43153	INTERCEPT (A) -	13.39858	SLOPE (B) -	0.79919
PLOTTED VALUES -	30	EXCLUDED VALUES-	0	MISSING VALUES -	0

'*********' IS PRINTED IF A COEFFICIENT CANNOT BE COMPUTED.

LANGUAGE GAIN ANALYSIS 08/11/81 PAGE 18

FILE NONAME (CREATION DATE = 08/11/81)

SCATTERGRAM OF (DOWN) C2 (ACROSS) A2



LANGUAGE GAIN ANALYSIS

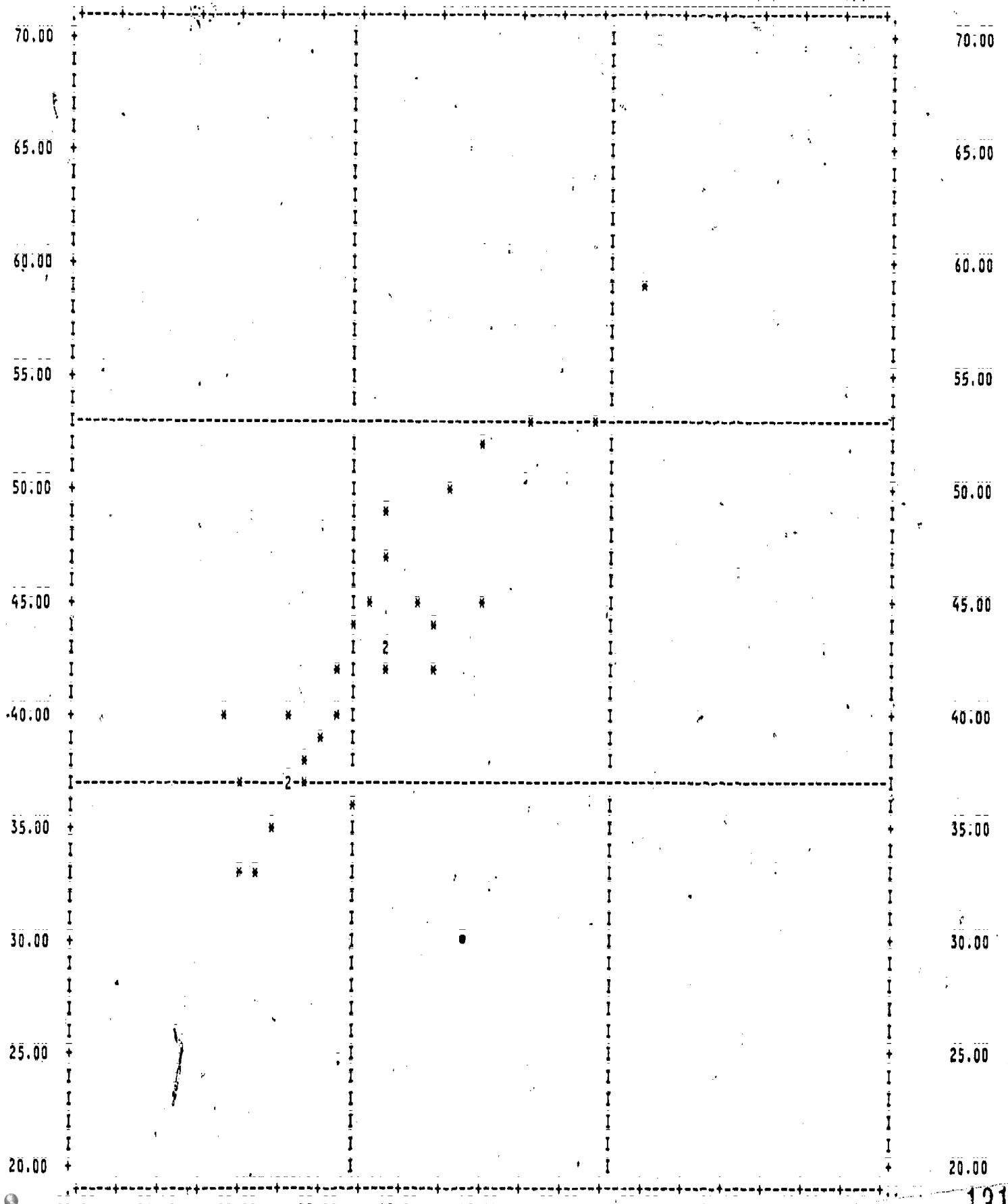
STATISTICS..

CORRELATION (R)-	0.73192	R SQUARED -	0.53571	SIGNIFICANCE -	0.00000
STD ERR OF EST -	4.15945	INTERCEPT (A) -	9.93491	SLOPE (B) -	0.87485
PLOTTED VALUES	30	EXCLUDED VALUES-	0	MISSING VALUES -	0

'*********' IS PRINTED IF A COEFFICIENT CANNOT BE COMPUTED.

LANGUAGE GAIN ANALYSIS 08/11/81 PAGE 23

FILE NONAME (CREATION DATE = 08/11/81)
SCATTERGRAM OF (DOWN) C3 (ACROSS) A3



STATISTICS..

CORRELATION (R)-	0.90148	R SQUARED -	0.81267	SIGNIFICANCE -	0.00000
STD ERR OF EST -	2.80576	INTERCEPT (A) -	8.17332	SLOPE (B) -	0.90297
PLOTTED VALUES -	30	EXCLUDED VALUES-	0	MISSING VALUES -	0

'*********' IS PRINTED IF A COEFFICIENT CANNOT BE COMPUTED.

```
SPSS BATCH SYSTEM                                                    08/17/81      PAGE    1

SPSS FOR OS/360, VERSION M, RELEASE 9.0, JUNE 10, 1981

                        CURRENT DOCUMENTATION FOR THE SPSS BATCH SYSTEM

ORDER FROM MCGRAW-HILL: SPSS, 2ND ED. (PRINCIPAL TEXT)      ORDER FROM SPSS INC.: SPSS STATISTICAL ALGORITHMS
                        SPSS UPDATE 7-9 (USE W/SPSS,2ND FOR REL. 7, 8, 9)          KEYWORDS: THE SPSS INC. NEWSLETTER
                        SPSS POCKET GUIDE, RELEASE 9
                        SPSS PRIMER (BRIEF INTRO TO SPSS)

DEFAULT SPACE ALLOCATION..     ALLOWS FOR..   102 TRANSFORMATIONS
WORKSPACE     71680 BYTES                     409 RECODE VALUES + LAG VARIABLES
TRANSPACE     10240 BYTES                    1641 IF/COMPUTE OPERATIONS

               1 RUN NAME      NEW VS. CONTINUING STUDENT TOEFL ANCOVA
               2 VARIABLE LIST A1,A2,A3,ATOT,RET,B1,B2,B3,BTOT,C1,C2,C3,CTOT
               3 INPUT MEDIUM  CARD
               4 INPUT FORMAT  FIXED(4X,3F2.0,1F3.0,65X,1F1.0/4X,3F2.0,1F3.0/4X,3F2.0,1F3.0)

                ACCORDING TO YOUR INPUT FORMAT, VARIABLES ARE TO BE READ AS FOLLOWS

                VARIABLE  FORMAT  RECORD     COLUMNS

                A1        F 2. 0      1      5-   6
                A2        F 2. 0      1      7-   8
                A3        F 2. 0      1      9-  10
                ATOT      F 3. 0      1     11-  13
                RET       F 1. 0      1     79-  79
                B1        F 2. 0      2      5-   6
                B2        F 2. 0      2      7-   8
                B3        F 2. 0      2      9-  10
                BTOT      F 3. 0      2     11-  13
                C1        F 2. 0      3      5-   6
                C2        F 2. 0      3      7-   8
                C3        F 2. 0      3      9-  10
                CTOT      F 3. 0      3     11-  13

THE INPUT FORMAT PROVIDES FOR  13 VARIABLES.  13 WILL BE READ
IT PROVIDES FOR   3 RECORDS ('CARDS') PER CASE.  A MAXIMUM OF   79 'COLUMNS' ARE USED ON A RECORD.

               5 N OF CASES    98
               6 ANOVA         CTOT BY RET(0,1) WITH ATOT
               7 STATISTICS    ALL

'ANOVA' PROBLEM REQUIRES     182 BYTES OF SPACE.

               8 READ INPUT DATA
```

NEW VS. CONTINUING STUDENT TOEFL ANCOVA

FILE NONAME (CREATION DATE = 08/17/81)

* * * * * * * * * * * * * CELL MEANS * * * * * * * * * * * * *

CIQT -
BY RET

* *

TOTAL POPULATION

455.34
(98)

RET

0	1
453.94	458.50
(68)	(30)

FILE NONAME (CREATION DATE = 08/17/81)

* * * * * * * * * A N A L Y S I S O F V A R I A N C E * * * * * * * * *

CTOT
BY RET
WITH ATOT

* *

SOURCE OF VARIATION	SUM OF SQUARES	DF	MEAN SQUARE	F	SIGNIF OF F
COVARIATES	260148.063	1	260148.063	283.878	0.000
ATOT	260148.063	1	260148.063	283.878	0.000
MAIN EFFECTS	1271.000	1	1271.000	1.387	0.242
RET	1271.007	1	1271.007	1.387	0.242
EXPLAINED	261419.063	2	130709.500	142.633	0.000
RESIDUAL	87058.688	95	916.407		
TOTAL	348477.750	97	3592.554		

COVARIATE	RAW REGRESSION COEFFICIENT
ATOT	0.841

98 CASES WERE PROCESSED.
0 CASES (0.0 PCT) WERE MISSING.

C-19

FILE NONAME (CREATION DATE = 08/17/81)

*** MULTIPLE CLASSIFICATION ANALYSIS ***

CTOT
BY RET
WITH ATOT

* *

GRAND MEAN = 455.34

VARIABLE + CATEGORY	N	UNADJUSTED DEV'N	UNADJUSTED ETA	ADJUSTED FOR INDEPENDENTS DEV'N	ADJUSTED FOR INDEPENDENTS BETA	ADJUSTED FOR INDEPENDENTS + COVARIATES DEV'N	ADJUSTED FOR INDEPENDENTS + COVARIATES BETA
RET							
0	68	-1.40				2.41	
1	30	3.16				-5.46	
			0.04				0.06
MULTIPLE R SQUARED							0.750
MULTIPLE R							0.866

C.20

CPU TIME REQUIRED.. 0.44 SECONDS

```
 9 ANOVA      CTOT BY RET(0,1) WITH ATOT,BTOT
10 STATISTICS ALL

'ANOVA' PROBLEM REQUIRES     266 BYTES OF SPACE.
```

NEW VS. CONTINUING STUDENT TOEFL ANCOVA 08/17/81 PAGE 6

FILE NONAME (CREATION DATE = 08/17/81)

* * * * * * * * * * * * * * * CELL MEANS * * * * * * * * * * * * * * *

CTOT
BY RET

* *

TOTAL POPULATION

455.34
(98)

RET

0	1
453.94	458.50
(68)	(30)

C.22

FILE NONAME (CREATION DATE = 08/17/81)

* * * * * * * * * A N A L Y S I S O F V A R I A N C E * * * * * * * * *

CTOT
BY RET
WITH ATOT
BTOT

* *

SOURCE OF VARIATION	SUM OF SQUARES	DF	MEAN SQUARE	F	SIGNIF OF F
COVARIATES	283109.875	2	141554.938	203.570	0.000
ATOT	3282.492	1	3282.492	4.721	0.032
BTOT	22961.793	1	22961.793	33.021	0.000
MAIN EFFECTS	3.750	1	3.750	0.005	0.942
RET	3.745	1	3.745	0.005	0.942
EXPLAINED	283113.625	3	94371.188	135.715	0.000
RESIDUAL	65364.125	94	695.363		
TOTAL	348477.750	97	3592.554		

COVARIATE	RAW REGRESSION COEFFICIENT
ATOT	0.245
BTOT	0.616

98 CASES WERE PROCESSED.
0 CASES (0.0 PCT) WERE MISSING.

C.23

NEW VS. CONTINUING STUDENT TOEFL ANCOVA

FILE NONAME (CREATION DATE = 08/17/81)

*** MULTIPLE CLASSIFICATION ANALYSIS ***

CTOT
BY RET
WITH ATOT
BTOT

* *

GRAND MEAN = 455.34

VARIABLE + CATEGORY	N	UNADJUSTED DEV'N	UNADJUSTED ETA	ADJUSTED FOR INDEPENDENTS DEV'N	ADJUSTED FOR INDEPENDENTS BETA	ADJUSTED FOR INDEPENDENTS + COVARIATES DEV'N	ADJUSTED FOR INDEPENDENTS + COVARIATES BETA
RET							
0	68	-1.40				0.13	
1	30	3.16				-0.30	
			0.04				0.00
MULTIPLE R SQUARED							0.812
MULTIPLE R							0.901

C.24

CPU TIME REQUIRED.. 0.28 SECONDS

```
11 ANOVA        CI BY RET(0,1) WITH AI
12 STATISTICS   ALL
```

'ANOVA' PROBLEM REQUIRES 182 BYTES OF SPACE.

NEW VS. CONTINUING STUDENT TOEFL ANCOVA

FILE NONAME (CREATION DATE = 08/17/81)

* * * * * * * * * * * * * CELL MEANS * * * * * * * * * * * * *

C1
BY RET

* *

TOTAL POPULATION

51.31
(98)

RET

0	1
51.29	51.33
(68)	(30)

FILE NONAME (CREATION DATE = 08/17/81)

* * * * * * * * * * * A N A L Y S I S O F V A R I A N C E * * * * * * * * * * *

C1
BY RET
WITH AI

* *

SOURCE OF VARIATION	SUM OF SQUARES	DF	MEAN SQUARE	F	SIGNIF OF F
COVARIATES	1985.384	1	1985.384	94.214	0.000
AI	1985.384	1	1985.384	94.214	0.000
MAIN EFFECTS	129.468	1	129.468	6.144	0.015
RET	129.468	1	129.468	6.144	0.015
EXPLAINED	2114.852	2	1057.426	50.179	0.000
RESIDUAL	2001.948	95	21.073		
TOTAL	4116.801	97	42.441		

COVARIATE	RAW REGRESSION COEFFICIENT
AI	0.650

98 CASES WERE PROCESSED.
0 CASES (0.0 PCT) WERE MISSING.

C.27

NEW VS. CONTINUING STUDENT TOEFL ANCOVA

FILE NONAME (CREATION DATE = 08/17/81)

*** MULTIPLE CLASSIFICATION ANALYSIS ***

C1
BY RET
WITH A1

****

GRAND MEAN = 51.31

VARIABLE + CATEGORY	N	UNADJUSTED DEV'N	UNADJUSTED ETA	ADJUSTED FOR INDEPENDENTS DEV'N	ADJUSTED FOR INDEPENDENTS BETA	ADJUSTED FOR INDEPENDENTS + COVARIATES DEV'N	ADJUSTED FOR INDEPENDENTS + COVARIATES BETA
RET							
0	68	-0.01				0.79	
1	30	0.03				-1.79	
			0.00				0.18
MULTIPLE R SQUARED							0.514
MULTIPLE R							0.717

C.28

CPU TIME REQUIRED.. 0.25 SECONDS

```
13 ANOVA        C1 BY RET(0,1) WITH A1,B1
14 STATISTICS   ALL
```

'ANOVA' PROBLEM REQUIRES 266 BYTES OF SPACE.

NEW VS. CONTINUING STUDENT TOEFL ANCOVA

FILE NONAME (CREATION DATE = 08/17/81)

* * * * * * * * * * * * * * CELL MEANS * * * * * * * * * * * * * *

C1
BY RET

* *

TOTAL POPULATION

51.31
(98)

RET

0	1
51.29	51.33
(68)	(30)

C.30

FILE NONAME (CREATION DATE = 08/17/81)

* * * * * * * * * * A N A L Y S I S O F V A R I A N C E * * * * * * * * * *

C1
BY RET
WITH A1
B1

* *

SOURCE OF VARIATION	SUM OF SQUARES	DF	MEAN SQUARE	F	SIGNIF OF F
COVARIATES	2727.258	2	1363.629	93.412	0.000
A1	29.455	1	29.455	2.018	0.159
B1	741.874	1	741.874	50.820	0.000
MAIN EFFECTS	17.336	1	17.336	1.188	0.279
RET	17.337	1	17.337	1.188	0.279
EXPLAINED	2744.594	3	914.865	62.671	0.000
RESIDUAL	1372.206	94	14.598		
TOTAL	4116.801	97	42.441		

COVARIATE	RAW REGRESSION COEFFICIENT
A1	0.130
B1	0.664

98 CASES WERE PROCESSED.
0 CASES (0.0 PCT) WERE MISSING.

NEW VS. CONTINUING STUDENT TOEFL ANCOVA

FILE NONAME (CREATION DATE = 08/17/81)

*** MULTIPLE CLASSIFICATION ANALYSIS ***

C1
BY RET
WITH A1
B1

* *

GRAND MEAN = 51.31

VARIABLE + CATEGORY	N	UNADJUSTED DEV'N	UNADJUSTED ETA	ADJUSTED FOR INDEPENDENTS DEV'N	ADJUSTED FOR INDEPENDENTS BETA	ADJUSTED FOR INDEPENDENTS + COVARIATES DEV'N	ADJUSTED FOR INDEPENDENTS + COVARIATES BETA
RET							
0	68	-0.01				0.30	
1	30	0.03				-0.68	
			0.00				0.07
MULTIPLE R SQUARED							0.667
MULTIPLE R							0.817

NEW VS. CONTINUING STUDENT TOEFL ANCOVA 08/17/81 PAGE 17

CPU TIME REQUIRED: 0.27 SECONDS

```
15 ANOVA        C2 BY RET(0,1) WITH A2
16 STATISTICS   ALL
```

'ANOVA' PROBLEM REQUIRES 182 BYTES OF SPACE.

NEW VS. CONTINUING STUDENT TOEFL ANCOVA

FILE NONAME (CREATION DATE = 08/17/81)

* * * * * * * * * * * * * C E L L M E A N S * * * * * * * * * * * * *
C2
BY RET
* *

TOTAL POPULATION

42.49
(98)

RET

0	1
42.46	42.57
(68)	(30)

FILE NONAME (CREATION DATE = 08/17/81)

* * * * * * * * * * A N A L Y S I S O F V A R I A N C E * * * * * * * * * *

C2
BY RET
WITH A2

* *

SOURCE OF VARIATION	SUM OF SQUARES	DF	MEAN SQUARE	F	SIGNIF OF F
COVARIATES	2889.447	1	2889.447	172.418	0.000
A2	2889.447	1	2889.447	172.418	0.000
MAIN EFFECTS	4.972	1	4.972	0.297	0.587
RET	4.972	1	4.972	0.297	0.587
EXPLAINED	2894.419	2	1447.209	86.357	0.000
RESIDUAL	1592.046	95	16.758		
TOTAL	4486.465	97	46.252		

COVARIATE	RAW REGRESSION COEFFICIENT
A2	0.742

98 CASES WERE PROCESSED.
0 CASES (0.0 PCT) WERE MISSING.

NEW VS. CONTINUING STUDENT TOEFL ANCOVA

FILE NONAME (CREATION DATE = 08/17/81)

*** MULTIPLE CLASSIFICATION ANALYSIS ***

C2
BY RET
WITH A2

* *

GRAND MEAN = 42.49

VARIABLE + CATEGORY	N	UNADJUSTED DEV'N	ETA	ADJUSTED FOR INDEPENDENTS DEV'N	BETA	ADJUSTED FOR INDEPENDENTS + COVARIATES DEV'N	BETA
RET							
0	68	-0.03				-0.15	
1	30	0.08				0.34	
			0.01				0.03
MULTIPLE R SQUARED							0.645
MULTIPLE R							0.803

C.36

CPU TIME REQUIRED.. 0.24 SECONDS

```
17 ANOVA      C2 BY RET(0,1) WITH  A2,B2
18 STATISTICS ALL
```

'ANOVA' PROBLEM REQUIRES 266 BYTES OF SPACE.

NEW VS. CONTINUING STUDENT, TOEFL ANCOVA

FILE NONAME (CREATION DATE = 08/17/81)

* * * * * * * * * * * * * CELL MEANS * * * * * * * * * * * * *

C2
BY RET

* *

TOTAL POPULATION

42.49
(98)

RET

0	1
42.46	42.57
(68)	(30)

C.38

FILE NONAME (CREATION DATE = 08/17/81)

* * * * * * * * * * A N A L Y S I S O F V A R I A N C E * * * * * * * * * *

C2
BY RET
WITH A2
B2

* *

SOURCE OF VARIATION	SUM OF SQUARES	DF	MEAN SQUARE	F	SIGNIF OF F
COVARIATES	3113.693	2	1556.846	106.664	0.000
A2	553.634	1	553.634	37.931	0.000
B2	224.246	1	224.246	15.364	0.000
MAIN EFFECTS	0.766	1	0.766	0.052	0.819
RET	0.766	1	0.766	0.052	0.819
EXPLAINED	3114.459	3	1038.153	71.127	0.000
RESIDUAL	1372.006	94	14.596		
TOTAL	4486.465	97	46.252		

COVARIATE	RAW REGRESSION COEFFICIENT
A2	0.500
B2	0.316

98 CASES WERE PROCESSED.
0 CASES (0.0 PCT) WERE MISSING.

C-39

NEW VS. CONTINUING STUDENT TOEFL ANCOVA

FILE NONAME (CREATION DATE = 08/17/81)

*** MULTIPLE CLASSIFICATION ANALYSIS ***

C2
BY RET
WITH A2
B2

****

GRAND MEAN = 42.49

VARIABLE + CATEGORY	N	UNADJUSTED DEV'N	ETA	ADJUSTED FOR INDEPENDENTS DEV'N	BETA	ADJUSTED FOR INDEPENDENTS + COVARIATES DEV'N	BETA
RET							
0	68	-0.03				-0.06	
1	30	0.08				0.13	
			0.01				0.01
MULTIPLE R SQUARED							0.694
MULTIPLE R							0.833

C.40

CPU TIME REQUIRED.. 0.27 SECONDS

```
19 ANOVA      C3 BY RET(0,1) WITH A3
20 STATISTICS ALL
```

'ANOVA' PROBLEM REQUIRES 182 BYTES OF SPACE.

C.41

NEW VS. CONTINUING STUDENT TOEFL ANCOVA

FILE NONAME (CREATION DATE = 08/17/81)

* * * * * * * * * * * * * * C E L L M E A N S * * * * * * * * * * * * * *

C3
By RET

* *

TOTAL POPULATION

42.52
(98)

RET

0	1
42.46	42.67
(68)	(30)

FILE NONAME (CREATION DATE = 08/17/81)

* * * * * * * * * A N A L Y S I S O F V A R I A N C E * * * * * * * * * *

C3
BY RET
WITH A3

* *

SOURCE OF VARIATION	SUM OF SQUARES	DF	MEAN SQUARE	F	SIGNIF OF F
COVARIATES	3348.730	1	3348.730	219.678	0.000
A3	3348.730	1	3348.730	219.678	0.000
MAIN EFFECTS	13.539	1	13.539	0.888	0.348
RET	13.539	1	13.539	0.888	0.348
EXPLAINED	3362.269	2	1681.135	110.283	0.000
RESIDUAL	1448.160	95	15.244		
TOTAL	4810.430	97	49.592		

COVARIATE	RAW REGRESSION COEFFICIENT
A3	0.889

98 CASES WERE PROCESSED.
0 CASES (0.0 PCT) WERE MISSING.

C-143

FILE NONAME (CREATION DATE = 08/17/81)

*** MULTIPLE CLASSIFICATION ANALYSIS ***

C3
BY RET
WITH A3

****

GRAND MEAN = 42.52

VARIABLE + CATEGORY	N	UNADJUSTED DEV'N	UNADJUSTED ETA	ADJUSTED FOR INDEPENDENTS DEV'N	ADJUSTED FOR INDEPENDENTS BETA	ADJUSTED FOR INDEPENDENTS + COVARIATES DEV'N	ADJUSTED FOR INDEPENDENTS + COVARIATES BETA
RET							
0	68	-0.06				0.25	
1	30	0.15				-0.56	
			0.01				0.05
MULTIPLE R SQUARED							0.699
MULTIPLE R							0.836

C-44

CPU TIME REQUIRED.. 0.24 SECONDS

```
21 ANOVA        C3 BY RET(0,1) WITH  A3,B3
22 STATISTICS   ALL
```

'ANOVA' PROBLEM REQUIRES 266 BYTES OF SPACE.

NEW VS. CONTINUING STUDENT TOEFL ANCOVA

FILE NONAME (CREATION DATE = 08/17/81)

* * * * * * * * * * * * * * C E L L M E A N S * * * * * * * * * * * * * *

C3
BY RET

* *

TOTAL POPULATION

42.52
(98)

RET

0	1
42.46	42.67
(68)	(30)

FILE NONAME (CREATION DATE = 08/17/81)

* * * * * * * * * A N A L Y S I S O F V A R I A N C E * * * * * * * * *

C3
BY RET
WITH A3
B3

* *

SOURCE OF VARIATION	SUM OF SQUARES	DF	MEAN SQUARE	F	SIGNIF OF F
COVARIATES	3920.842	2	1960.421	207.353	0.000
A3	84.767	1	84.767	8.966	0.004
B3	572.112	1	572.112	60.512	0.000
MAIN EFFECTS	0.865	1	0.865	0.092	0.763
RET	0.865	1	0.865	0.092	0.763
EXPLAINED	3921.707	3	1307.236	138.266	0.000
RESIDUAL	888.722	94	9.454		
TOTAL	4810.430	97	49.592		

C-47

COVARIATE	RAW REGRESSION COEFFICIENT
A3	0.275
B3	0.653

98 CASES WERE PROCESSED.
0 CASES (0.0 PCT) WERE MISSING.

NEW VS. CONTINUING STUDENT TOEFL ANCOVA

FILE NONAME (CREATION DATE = 08/17/81)

*** MULTIPLE CLASSIFICATION ANALYSIS ***

C3
BY RET
WITH A3
B3

GRAND MEAN = 42.52

VARIABLE + CATEGORY	N	UNADJUSTED DEV'N	UNADJUSTED ETA	ADJUSTED FOR INDEPENDENTS DEV'N	ADJUSTED FOR INDEPENDENTS BETA	ADJUSTED FOR INDEPENDENTS + COVARIATES DEV'N	ADJUSTED FOR INDEPENDENTS + COVARIATES BETA
RET							
0	68	-0.06				-0.06	
1	30	0.15				0.14	
			0.01				0.01
MULTIPLE R SQUARED							0.815
MULTIPLE R							0.903

C.48

CPU TIME REQUIRED.. 0.28 SECONDS

23 FINISH

NORMAL END OF JOB.
23 CONTROL CARDS WERE PROCESSED.
0 ERRORS WERE DETECTED.

TOEFL Research Reports currently available . . .

Report 1. *The Performance of Native Speakers of English on the Test of English as a Foreign Language.* John L. D. Clark. November 1977.

Report 2. *An Evaluation of Alternative Item Formats for Testing English as a Foreign Language.* Lewis W. Pike. June 1979.

Report 3. *The Performance of Non-Native Speakers of English on TOEFL and Verbal Aptitude Tests.* Paul J. Angelis, Spencer S. Swinton, and William R. Cowell. October 1979.

Report 4. *An Exploration of Speaking Proficiency Measures in the TOEFL Context.* John L. D. Clark and Spencer S. Swinton. October 1979.

Report 5. *The Relationship between Scores on the Graduate Management Admission Test and the Test of English as a Foreign Language.* Donald E. Powers. December 1980.

Report 6. *Factor Analysis of the Test of English as a Foreign Language for Several Language Groups.* Donald E. Powers and Spencer S. Swinton. December 1980.

Report 7. *The Test of Spoken English as a Measure of Communicative Ability in English-Medium Instructional Settings.* John L. D. Clark and Spencer S. Swinton. December 1980.

Report 8. *Effects of Item Disclosure on TOEFL Performance.* Gordon A. Hale, Paul J. Angelis, and Lawrence A. Thibodeau. December 1980.

Report 9. *Item Performance Across Native Language Groups on the Test of English as a Foreign Language.* Donald L. Alderman and Paul W. Holland. August 1981.

Report 10. *Language Proficiency as a Moderator Variable in Testing Academic Aptitude.* Donald L. Alderman. November 1981.

Report 11. *A Comparative Analysis of TOEFL Examinee Characteristics, 1977-1979.* Kenneth M. Wilson. September 1982.

Report 12. *GMAT and GRE Aptitude-Test Performance in Relation to Primary Language and Scores on TOEFL.* Kenneth M. Wilson. October 1982.

Report 13. *The Test of Spoken English as a Measure of Communicative Ability in the Health Professions: Validation and Standard Setting.* Donald E. Powers and Charles W. Stansfield. January 1983.

If you wish additional information about TOEFL research or would like to be placed on the mailing list to automatically receive order forms for newly published reports, write to:

TOEFL Program Office
Educational Testing Service
Princeton, NJ 08541
USA

5790606. S5] 12. 275569. Printed in U.S.A.